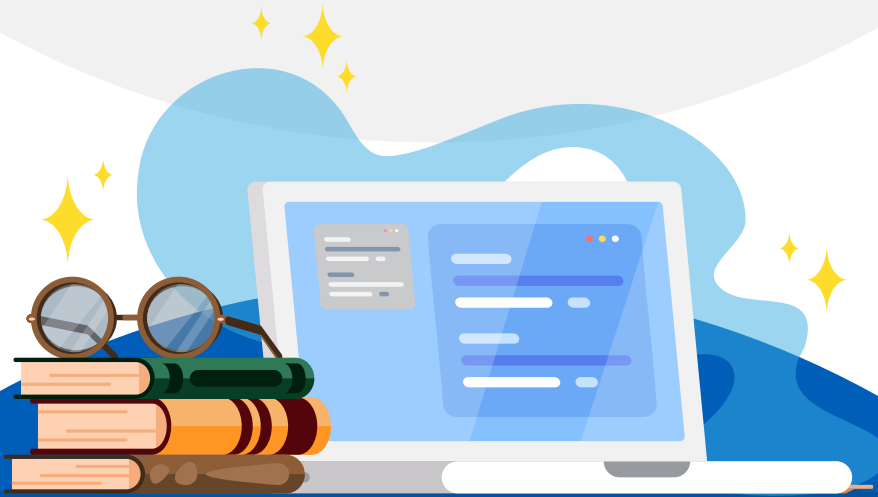


Regulating Harmful Content in Indonesia:

Legal Frameworks, Trends, and Concerns





Authors



Faiz Rahman
Sri Handayani Nasution
Aridiva Firdharizki
Nadya Olga Aletha
Alfredo Putrawidjoyo

Reviewer



Novi Kurnia

Designer



Riawan Hanif Alifadecya

©2022 Center for Digital Society, Universitas Gadjah Mada, All Rights Reserved.

No part of this publication may be reproduced or transmitted, in any form or by any means (electronic, mechanical, photocopying, recording or otherwise), without prior permission of both the copyright owner and the publisher of the book.

Disclaimer

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

The ideas and opinions expressed in this document are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.

This document was produced with the financial support of the European Union. Its contents are the sole responsibility of the authors and do not necessarily reflect the views of the European Union.



Funded by
the European Union



Content

I	Content
ii	List of Tables
ii	List of Figures
iii	Executive Summary
ix	Glossary
1	Introduction
2	• Background
7	• Methodology
10	Literature Review
11	• Defining Illegal and Harmful Content through International and Regional Legal Instruments
18	• International Norms and Standards on Regulating Illegal and Harmful Content
25	• Platforms and Content Moderation
28	Chapter I :
29	Regulations of Online Harmful Content in Indonesia
32	• Constitutional Basis on Regulating 'Illegal and Harmful Content' in Indonesia
51	• Classification of Illegal and Harmful Content in Indonesian Regulations
54	• Applicable Remedies and Handling Method of Illegal and Harmful Content based on Indonesian Regulations
	• The Responsibilities of Social Media Platforms in Regulating Illegal and Harmful Content

58	Chapter II : Trends
59	<ul style="list-style-type: none"> • Trends through the Lens of State and Government Institutions: Policies Issued in Responding Content-Related Problems in Societies
66	<ul style="list-style-type: none"> • Trends Through the Lens of Societies
71	<ul style="list-style-type: none"> • Tech-based and Automated Content Moderation
75	Chapter III : Concern
76	<ul style="list-style-type: none"> • Regulating Grey Area: Hate Speech, Misinformation and Disinformation, and Defamation
78	<ul style="list-style-type: none"> • The Gap between Platforms' Self-Regulatory Mechanism and Regulations
84	<ul style="list-style-type: none"> • Impact of Content Regulations Toward Societies
97	Chapter IV : Conclusions and Recommendations
98	<ul style="list-style-type: none"> • Conclusions
102	<ul style="list-style-type: none"> • Recommendations
106	References

→ List of Tables

46	Table 1.	Online Content Classification Based on Indonesian Regulations
80	Table 2.	Comparison of Content Classification in Indonesia's Regulation and Platforms' Community Guidelines
84	Table 3.	Comparison between State and Social Media Platforms' Approach in Handling Illegal and Harmful Content in Indonesia

→ List of Figures

20	Figure 1.	Six-Part Thresholds Test in Applying Article 20(2) of the ICCPR
----	------------------	---



Executive Summary

This research is part of the Social Media 4 Peace (SM4P) project, conducted by the Center for Digital Society, Faculty of Social and Political Science, Universitas Gadjah Mada, in a partnership with the United Nations Educational, Scientific and Cultural Organization (UNESCO) with financial support from the European Union (EU). As the first phase of the overall project, this research aims to enhance the understanding of legal frameworks, trends, and concerns regarding harmful content regulation and its implementation in Indonesia to further strengthen society's resilience from potentially harmful content while also ensuring the protection of their freedom of expression and speech in digital space.

Content-sharing features in User Generated Content (UGC) internet platforms are often being misused to transmit types of content that might be harmful or even illegal. In Indonesia, dis/misinformation and online hate speech issues—the focus of this research—have become more pressing. Back in 2019, disinformation, circulated on social media during the presidential election, contributed to fuelling a violent riot. Cases of hate speech online against the LGBTQ+ community and religious minority regularly occur. Both social media platforms and governments have attempted to minimize the harm users experience while using the Internet. However, these attempts are not without flaws and/or repercussions on freedom of expression.

Indonesia already has several laws and regulations for harmful content, such as mis/disinformation and hate speech, including: Law No. 11 of 2008 on Electronic Information and Transaction and its 2016 revision (the EIT Act) and its implementing regulations; the Government Regulation No. 71 of 2019 on Electronic System and Transaction Implementation (GR ESTI); and the Regulation of Minister of Communication and Informatics No. 5 of 2020 (MOCI Regulation 5/20). However, these laws and regulations face some

criticism from civil organizations, activists, and academics. Some provisions under the existing legal frameworks on content regulations may be prone to misuse by the government—misuse that may lead to violations of freedom of expression.

Key Findings

→ Legal Framework

Publication and dissemination of harmful content are constituted as criminal offenses according to Indonesian regulation. Consequently, all harmful content could be criminally prosecuted

There is no distinction between illegal and harmful content across content regulations in Indonesia. Types of content listed in the available regulations are all treated as illegal and, therefore, as criminal offenses, including content that contains disinformation and hate speech. As such, there are only two applicable mechanisms in handling harmful content under Indonesia's regulation: court mechanisms (criminal prosecution, and for some instances, civil lawsuits), and non-court mechanisms (alternative dispute resolution and administrative actions).

Several terms used in the regulations are too broad (e.g., morality, public order, etc.) and may cause multiple interpretations, misinterpretations, and controversy

Although many policies related to the handling of harmful online content have been issued by state and government institutions, the implementation of the EIT Act still causes multiple interpretations and controversy among civil society. This is demonstrated by various content-related cases handled by the police and discrepancies in judicial interpretation of terminologies related to harmful content in many court decisions.

Indonesia is adopting a punitive approach of online content regulation

Several regulations of content also govern the social media platforms—legally termed as the Electronic System Operators (ESO). These regulations are the EIT Act, GR ESTI, and the MOCI Regulation 5/20. These regulations offer the classification of content along with several rules that platforms need to comply with, including responding to take down requests within a limited timeframe. The regulation does not provide any due process on take-down requests, especially those made by the government. The short timeframe does not give the ESOs sufficient time to assess the content removal request carefully. Ultimately, it may force the ESOs to comply to avoid administrative punishment—the regulations are therefore adopting a punitive approach in their implementation and design.

→ Trends and Concerns

Disparity in harmful content regulation between government regulations and platforms' self-regulatory mechanisms

In Indonesia, the disparity start from the classification of harmful content. This disparity in classifying content leads to differences in the handling mechanism. On social media platforms, the platforms will only remove harmful content that violates community guidelines. Still, they mostly resort to other—arguably softer—means of moderation when content is not violating the guidelines but may be considered harmful regardless. For instance, platforms may resort to flagging, labeling, downranking, and demonetizing harmful content. In contrast, Indonesia treats all harmful content as illegal. Therefore, content removal is the only method legally recognized to moderate 'illegal' content in Indonesia based on the existing regulations.

Neglect and transparency

There are allegations that platforms' investments in moderating content in non-English languages are severely underfunded. There is a

considerable lack of information on how platforms are employing local content moderators or how they are conducting content moderation practices in Indonesia. On the other hand, in its tech-based—or automated—content moderation that uses algorithms may not be as effective as the platforms claim. For one, automated content moderation may be biased, rooted from the biases made by human moderators. Hence, automated content moderation may make ineffective—even harmful—decisions if deployed without sensitive consideration of social, cultural, and political divergences worldwide. Unfortunately, there is not much information available on how platforms employ their local content moderator in Indonesia nor how they teach their algorithms—for automated content moderation. Thus platforms’ practices in Indonesia generally lack transparency.

A lack of transparency is also apparent on the government’s side. For instance, the information on the government’s content removal requests to platforms is not publicly available. They also do not provide appeal mechanisms for their requests.

Regulations on content disproportionately affect the marginalized community

The EIT Law and its implementation are frequently on the side of violating freedom of expression, which is far from ensuring peace and national stability. The law is posited to give too much power and discretion to law enforcement without due process and robust accountability measures. Several groups of people often become the target of this regulation, namely journalists and civil societies, the Ahmadiyya community, gender and sexual minorities, and ordinary citizens. For some marginalized groups like the Ahmadis and the LGBTQ+ community, these regulations may severely violate their rights, on top of the already existing structural and systematic violence they face on a daily basis.

Key Recommendations

As there is still room for improvement in content moderation, strong commitment both from the government and the private sector, such as social media platforms, is critically needed. Several recommendations based on our analysis are:

Recommendation 1: Revising the EIT Act and its implementing regulations

The EIT Act serves as the primary legal basis for regulating cyberspace in Indonesia. However, several things need to be improved in its implementation, both in the EIT Act and its implementing regulations. The government needs to: reconsider the harmful content classification provisions so that not all of them can be categorised as criminal acts; redefine terms related to 'illegal' and 'harmful' content; and reform the existing handling mechanism for harmful content, including reformulating sanctions and developing comprehensive approaches toward harmful content through education and technological means.

Recommendation 2: Harmonizing the laws and regulations related to illegal and harmful content

The harmonization is projected to reduce the possibility of different interpretations and overlaps between regulations. This recommendation is not only carried out on the regulations governing the online realm, but also on other regulations that may intersect with the issue of illegal and harmful content.

Recommendation 3: Equalizing perceptions of the meaning of the provisions of actions prohibited in the EIT Act

This research recommends the State to provide common perceptions of the meaning of the provisions of actions prohibited in the EIT Act. In

some cases, an act can be interpreted differently by the police, prosecutor, and judiciary. Therefore, there must be a unified perception and interpretation of the provisions of prohibited acts. This can be achieved by making a joint decree or other legal instruments. Furthermore, to determine whether the content is harmful or not, the police, prosecutor, and judiciary could adopt the Rabat Plan of Action and its six-art threshold test, which includes: context, speaker, intent, content and form, the extent of the speech act, and likelihood. Additionally, training for police, prosecutor, and the judiciary should also be carried out to gain a shared understanding of applying provisions related to illegal and harmful content.

Recommendation 4: Enhancing cooperation between State and social media platforms and relevant local stakeholders in handling illegal and harmful content

The State and social media platforms have shared responsibility for handling harmful online content. Thus, it must be ensured that the state and social media platforms are moving in the same direction in dealing with illegal and harmful online content. Therefore, various discussions and multi-stakeholder meetings need to be encouraged.

Recommendation 5: Increasing transparency in moderating content

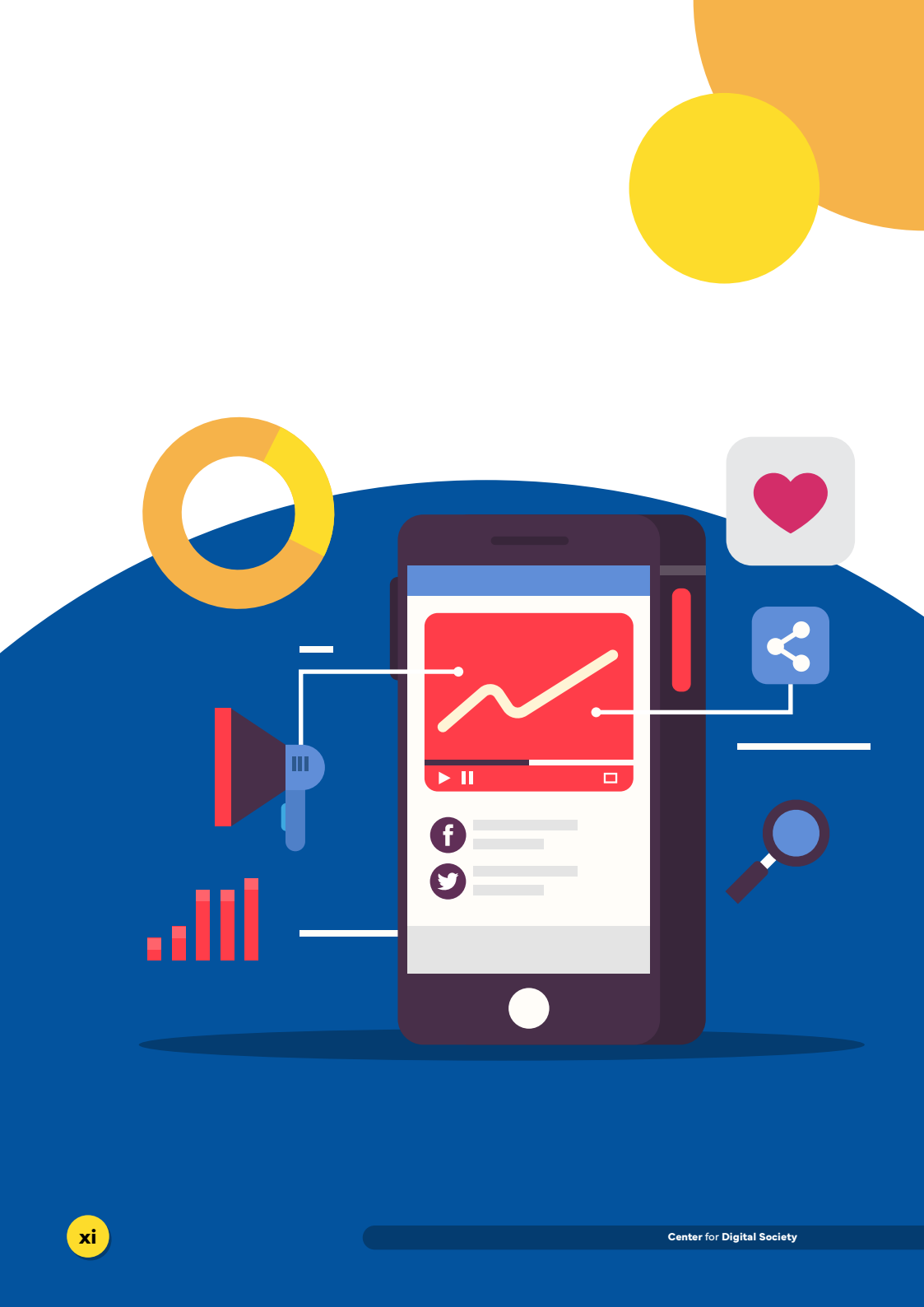
Both the State and social media platforms need to ensure there is meaningful transparency in the implementation of their content regulation. Meaningful transparency means government or social media platforms do not merely produce an output-based report (e.g., numbers of content being moderated) but include the information on the whole process of content moderation (selection criteria of local content moderators, transparency on classification of the 'harmful content' and the underlying process of how content are moderated).

→ Glossary

AI	Artificial Intelligence
AMRI	ASEAN Ministers Responsible for Information
APC	Association for Progressive Communications
ARED	Abolition of Race and Ethnic Discrimination
ASEAN	Association of Southeast Asian Nations
CSIS	Center for Strategic and International Studies
COVID-19	Coronavirus Disease 2019
DSA	Digital Services Act
ED	Electronic Document
EDI	Electronic Data Interchange
ESO	Electronic System Operator
ESTI	Electronic System and Transaction Implementation
EI	Electronic Information
EIT	Electronic Information and Transaction
EU	European Union
GR	Government Regulation
HRC	Human Rights Committee
ICCPR	the International Covenant on Civil and Political Rights
ICERD	the International Convention on the Elimination of All Forms of Racial Discrimination
ICESCR	the International Covenant on Economic, Social and Cultural Rights
ICT	Information and Communications Technology
IHRL	International Human Rights Law
KPU	Komisi Pemilihan Umum (General Election Commission)

LGBTQ+	Lesbian, Gay, Bisexual, Transgender, Intersex, Queer/Questioning, Asexual and other sexual preferences, and gender identities.
MK	Mahkamah Konstitusi (Constitutional Court of the Republic of Indonesia)
MOCI	Minister of Communication and Informatics of the Republic of Indonesia
OHCHR	Office of the High Commissioner for Human Rights
PN	Pengadilan Negeri (District Court)
SAFEnet	Southeast Asia Freedom of Expression Network
SM4P	Social Media 4 Peace
The 1945 Constitution	The 1945 Constitution of the Republic of Indonesia
TRIP	Trade-Related Aspects of Intellectual Property
UDHR	Universal Declaration of Human Rights
UN	United Nations
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
USAID	United States Agency for International Development
UGC	User Generated Content
WEF	World Economic Forum







Introduction

Background

As of 2021, the number of social media users reached 4.62 billion globally, experiencing 10.1% more growth than the previous year (Hootsuite, 2022). In Indonesia, by February 2022, the number of active social media users exceeded 191.4 million, or approximately 68.9% of the population (Hootsuite, 2022). The number increased by around 12.6% compared to last year. The most-used social media platforms in Indonesia are WhatsApp, Instagram, and Facebook, with 88.7%, 84.8%, and 81.3% monthly internet users (Hootsuite, 2022). With the increasing number of social media users in Indonesia, the potential for users to be exposed to illegal and harmful content is also increasing.

Within social media platforms, users are able to produce and share content in the form of images, videos, comments, products, advertisements, and other types of content. This content is usually known as User Generated Content (UGC) (Setiawan, 2021). This content-sharing feature facilitates users to express ideas, exchange information, and interact. Unfortunately, this feature is also often used to transmit types of content that might be harmful or even illegal, such as content containing terrorism, sexual exploitation, hate speech, and misinformation or disinformation.

In 2019, the Ministry of Communication and Informatics (MOCI) of Indonesia received more than 430 thousand reports of harmful content, with pornography topping the chart, followed by fitnah (defamation), disturbing content, gambling, scams, mis/disinformation (hoaxes), and many others (MOCI, 2020). Among these varieties of harmful content, mis/disinformation and hate speech—the focus of this research—are becoming more pressing in Indonesia.

This research outlines the understanding of misinformation and disinformation as follows: both misinformation and disinformation are classified as the act of spreading false information, though with different intentions. For misinformation, the person who spreads it usually does not

understand that the information is untrue, meanwhile for disinformation, the person deliberately spreads false information with ill intent to cause people being actively disinformed (UNESCO, 2018). Hate speech is a public, ill-intentioned act of speech targeted against members of systematically marginalised groups to inflict a sense of inferiority, legitimate discriminatory behaviour, and deprive the targeted groups of power (Gelber, 2019, retrieved from Sinpeng et al., 2021). In addition, this research also includes discussions on defamation in Indonesia in this report. While the understanding of defamation is vague, this report finds the urgency to discuss how the government regulates the types of content they understand as 'defamation'. This research sees how the effort to regulate this issue may facilitate more restrictions on freedom of speech in Indonesia.

MOCI has been concerned about mis/disinformation cases for several years. The government body regularly updates the mis/disinformation cases through their website.¹ The increasing number of mis/disinformation cases can be seen, for instance, during the election period (Katadata, 2020). During the 2019 national election, ethnic, and religiously-tinged mis/disinformation contributed to fuelling violet riots (Temby, 2019).

Content containing hate speech also shows a significant increase. Since 2018, MOCI has handled and taken down 3640 cases of online hate speech content based on ethnicity, religion, race, and intergroups (MOCI, 2021). According to the Center for Strategic and International Studies (CSIS), the phenomenon of hate speech in Indonesia has experienced a rapid increase over the past decade (CSIS, 2021). In particular, cases of online hate speech against the LGBTQ+ community and religious minorities are reported to occur more frequently (Sinpeng et al., 2021; Lina et al., 2021). Based on the CSIS report, the number of cases of hate speech content directed towards vulnerable communities like Ahmadiyya, Shi'a, and Chinese Indonesians increased significantly in 2021.²

¹ See the MOCI Public Information regarding mis/disinformation cases through https://eppid.kominfo.go.id/informasi_publik/Informasi%20Publik%20Setiap%20Saat.

² See CSIS report on the trends of hate speech cases toward vulnerable communities in <https://dashboard.csis.or.id/hatespeech/#trends>

On the other hand, aside from mis/disinformation and hate speech cases, defamation content topped the reported cases in 2020 with 1743 out of 4656 reports (CNN, 2020). The report from SAFEnet shows similar findings and states that the number of convictions related to harmful online content has quadrupled compared to the previous year and is mainly based on defamation, hate speech, and mis/disinformation (SAFEnet, 2021). Nevertheless, mis/disinformation, hate speech, and defamation cases are interrelated in many instances, as many hate speech or defamation cases are caused by mis/disinformation content.³

Indonesia already has several laws and regulations for illegal and harmful content, including, Law No. 11 of 2008 on Electronic Information and Transaction (EIT Act), and its implementing regulations, such as Government Regulation No. 71 of 2019 on Electronic System and Transaction Implementation (GR ESTI). However, these laws and regulations face criticism from civil organizations, activists, and academics. Ever since the stipulation of the EIT Act in 2008, various organizations, activists, and scholars have criticized the EIT Act's implementation, as it is argued to threaten freedom of speech (SAFEnet, 2021; Amnesty International Indonesia et al., 2021). Most of the criticism of the EIT Act revolves around the formulation and interpretation of several 'prohibited acts', including defamation, misinformation on electronic transactions, hate speech, and interception. Additionally, since 2008, the constitutionality of these provisions has been tested several times through submission for constitutional review to the Constitutional Court.⁴ Interestingly, in all decisions reviewing 'prohibited acts', only one was granted by the Constitutional Court regarding illegal interception. However, in all decisions concerning illegal and harmful content, such as defamation and hate speech, the Court held that these provisions are constitutional, although the reality on the ground shows otherwise.

³ For instance, Lucky Alamsyah in 2021 was reported to the police by Roy Suryo on the basis of defamation because Alamsyah allegedly share fake news regarding Roy Suryo. See CNN Indonesia 'Roy Suryo Polisikan Lucky Alamsyah Kasus Pencemaran Nama Baik' <https://www.cnnindonesia.com/nasional/20210524175343-12-646388/roy-suryo-polisikan-lucky-alamsyah-kasus-pencemaran-nama-baik>.

⁴ See e.g., Constitutional Court Decision No. 50/PUU-VI/2008 (defamation case), Constitutional Court Decision No. 2/PUU-VII/2009 (defamation case), Constitutional Court Decision No. 52/PUU-XI/2013 (hate speech case), Constitutional Court Decision No. 76/PUU-XV/2017 (hate speech case).

In fact, freedom of speech and expression safeguarded by the International Covenant on Civil and Political Rights has been guaranteed in the 1945 Constitution of the Republic of Indonesia (the 1945 Constitution) and the Indonesian Human Rights Act (Law No. 39 of 1999). The post-amendment 1945 Constitution provides a specific chapter regarding human rights, which also includes the protection of freedom of speech and expression. It is further emphasized in the Human Rights Act (Law No. 39 of 1999) through several Articles, including Article 14, Article 23 paragraph (2), and Article 25. Moreover, Article 28J paragraph (2) of the 1945 Constitution also emphasizes that the limitation of such rights should only be based on the Acts and considering the just demands based on morality, religious values, security, and public order in a democratic society.

Despite the existence of a significant constitutional and legal basis for freedom of speech and expression in Indonesia, the implementation in practice remains a problem, especially in this digital age. Under the following legal framework that becomes the basis for regulating content—such as the EIT Act, the Criminal Code etc.—the definition (e.g., morality, public order, etc.) that limits freedom of expression is too vague and potentially misinterpreted. This could lead to misuse by law enforcement officials during a criminal investigation or judicial proceedings. Furthermore, it could also lead to abuse by the government and/or the authorities, such as silencing criticism toward government policies for the aforementioned political purposes.

Moreover, within democratic countries, respecting the right of each other and public order becomes the priority issue in the implementation of the right to freedom of speech (Anindyajati, 2021). As freedom of expression is recognized globally, it is also essential to investigate the relevant international legal instruments, including the international standards of illegal and harmful content, which revolve around the freedom of expression issues, especially in the digital space.

As part of the affected parties from the existing illegal and harmful content regulation in the digital space, unravelling each social media platform's content policies is also necessary. However, there is an incongruence between the existing regulation and social media platforms' guidelines (Haryanto, 2020). Take, for example, both parties' definitions of 'restricted content'. The 'restricted content' is not explicitly mentioned in the EIT Act but could be interpreted as part of 'Prohibited Acts' regulated in Chapter VII. The 'Prohibited Acts' related to online content include content that contains gambling, moral code violations, blasphemy and/or defamation, extortion and/or threat, and hate speech and false news in electronic transactions.

On the other hand, social media platforms define 'restricted content' as 'harmful content' and translate it to content that has to either be removed or content that has to be limited. While the must-remove content aligns with the Indonesian government's definition, the latter can differ. In cases of false news, social media only resorts to limiting its exposure rather than completely removing the content from the platforms since it is difficult to differentiate between false news and opinion.

To summarise, it is essential to reformulate the existing Indonesian digital content regulations to ensure the protection of freedom of speech and expression. Furthermore, the firm and robust protection of freedom of expression also constitutes an essential foundation for democracy, the rule of law, peace, stability, sustainability, inclusive development, and participation in public affairs (the Council of the European Union, 2014).

This research is the first phase of the overall Social Media 4 Peace (SM4P) project, conducted by the Center for Digital Society UGM, in partnership with UNESCO and funded by the EU. The objective of this research is to systematically map the national laws, regulations, and policies related to illegal and harmful content and analyse the issues arising from the existing content regulation in Indonesia's legal framework. This research also analyse the current international legal framework and social media platform policies in relation to illegal and harmful content.

→ Methodology

This research is carried out by conducting a literature review and regulation mapping and analysis to understand how existing legal frameworks regulate online content in Indonesia, how the regulations are implemented, what their impact is on the freedom of speech in Indonesia, and how they affect the least-protected communities, such as the LGBTQ+ community and religious minorities online.

→ Literature Review

The literature review acts as a baseline for this study; therefore, it is conducted to understand several key points:

● Terminology of Illegal and Harmful Content

- Examining the development of illegal and harmful content definition and its relevant norms on an international and regional level.
- Determining the use of 'harmful' and 'illegal' content in content moderation, and inquiring whether the differences in these terms affect the implementation of content moderation regulations and mechanisms.

● Existing Norms and Practices on Content Moderation

Understanding the internationally recognized best practices of content moderation and the norms set by various international or regional organizations.

● Platforms and Content Moderation

Inquiring into the roles of private entities in eliminating the transmission of illegal and harmful content within their platforms and reviewing the performance and transparency in their implemented self-regulatory mechanisms.

Regulation Mapping and Analysis of Indonesian

→ Regulations Related to 'Harmful Content' and its Implementation

The regulation mapping comprises two analyses: regulations and policies mapping, and implementation case analyses. Regulations and policies mapping is conducted to explore several keypoints:

- how the government defines and characterizes illegal or harmful content;
- the applicable remedies and handling method of harmful content;
- the responsibilities of social media platforms;
- self-regulatory mechanism initiated by social media platforms.

The regulations mapping and analysis were conducted in four stages:

● **Identify and analyse the primary and related regulations concerning illegal and harmful online content.**

The EIT Act and its implementing regulations, as it is mainly used to handle online-related activities, including illegal and harmful content, are analyzed. Furthermore, relevant existing regulations that could be used in determining and handling harmful online content are examined based on the types of illegal and harmful content mentioned in the EIT Act and its implementing regulations.

● **Identify and analyse court decisions related to illegal and harmful online content.**

Based on the identified regulations, court decisions (both in the Constitutional Court and Supreme Court) are examined to support the regulatory analysis, especially how the judiciary interprets and implements the regulations through court proceedings.

Identify and analyse the relevant policies issued by the government and state institutions.

Furthermore, through analysis of current regulations, policies issued by the government or state institutions to respond to public demand concerning online illegal and harmful content cases are also identified.

Analysis of the implementation of state regulations and policies toward society

This part focuses on how the existing regulations may affect the community, with the presumption that the current law enforcement on online content may compromise the freedom of expression and disproportionately affect the least protected communities such as the LGBTQ+ and religious minorities.



Literature Review

The background features several overlapping circular and semi-circular shapes. A large, light blue shape is on the right side. A darker blue shape is at the bottom. An orange circle is on the left side, partially overlapping the darker blue shape.

The literature review will determine the current practices of content moderation: for what purposes it is mainly used, by whom, and on which platforms. Importantly, this part of the study helps identify the prominent issues in content moderation as a whole, contextualizing it in the current pressing issue of the spread of misinformation, disinformation, and hate speech. Additionally, harmful and/or illegal content moderation practices are presumably being used at the expense of the marginalized communities, resulting in various forms of unintended and intended censorship that may disproportionately impact the least protected communities. The literature review will be as follows: assessing the international and regional norms on illegal and harmful content and the roles of platforms in moderating content.

Defining Illegal and Harmful Content through International and Regional Legal Instruments

→ Global Legal Instruments

One of the obstacles in effectively regulating content online is harmonizing the definition and/or classification of what constitutes 'harmful' and 'illegal' content. Unfortunately, no treaties in the international legal regime specifically regulate illegal and harmful content on the Internet. The International Human Rights Law (IHRL) only implicitly regulate illegal content through broad concepts like racial discrimination. Therefore, the distinction between the two terms remains unexplored in the available international treaties. Nevertheless, the **Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (A/66/290)** which was submitted to the General Assembly at its sixty-sixth session on freedom of expression on the Internet, has presented a discussion on the distinction between illegal and harmful content. The Special Rapporteur underscores the statement as follows:

*'... there are differences between **illegal** content, which States are required to prohibit under international law, such as child pornography, and those that are considered **harmful, offensive, objectionable or undesirable**, but which States are neither required to prohibit nor criminalize. In this regard, the Special Rapporteur believes that it is important to make a clear distinction between three types of expression: (a) **expression that constitutes an offense** under international law and can be prosecuted criminally; (b) **expression that is not criminally punishable** but may justify a restriction and a civil suit; and (c) **expression that does not give rise to criminal or civil sanctions**, but still raises concerns in terms of tolerance, civility, and respect for others. **These different categories of content pose different issues of principle and call for different legal and technological responses.**'*

The Special Rapporteur then provides four types of expression, which **fall under the first category**, which is the expression that constitutes an offense under international law and can be prosecuted criminally. The four types of expressions include: (1) child pornography; (2) direct and public incitement to commit genocide; (3) advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence; and (4) incitement to terrorism. With a note, they must also comply with the three-part test of prescription by unambiguous law, pursuance of a legitimate purpose, and respect for the principles of necessity and proportionality.

The Special Rapporteur further mentioned that types of expression that **do not fall under the first category** should not be criminalized, including defamation laws aimed at protecting the reputation of individuals. As stipulated in the Human Rights Council Resolution 12/16, the Special

Rapporteur stresses that the following types of expression **should never be subject to restrictions**: discussion of government policies and political debate; reporting on human rights, government activities, and corruption in government; engaging in election campaigns, peaceful demonstrations or political activities, including for peace or democracy; and expression of opinion and dissent, religion or belief, including by persons belonging to minorities or vulnerable groups.

Moreover, several types of harmful content in general still lack clarity in their definition. It is feared that it could endanger people's rights to freedom and expression. To anticipate this, several UN instruments have provided definitions of various harmful content, such as hate speech, disinformation, and misinformation—which are the focus of this research.

The United Nations Strategy and Plan of Action on Hate Speech defined hate speech as speech that attacks or uses pejorative or discriminatory language with reference to a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factors. Even though it has been defined in such a way, regulating hate speech must also meet the criteria stated in the ICCPR. These criteria include several points, which will be discussed further in the next section.

Concerning disinformation and misinformation, the Special Rapporteur (A/HRC/47/25) clarified both terms' definitions. Disinformation is defined as false information that is disseminated intentionally to cause serious social harm, while misinformation is defined as the dissemination of false information unknowingly. Furthermore, the instrument emphasized that both terms are not interchangeable.

When it comes to defamation, the existing legal instruments do not currently clarify the definition of the term. However, defamation laws are usually justified to protect the reputation of individuals. Several Special Rapporteur instruments have now called on the state to repeal all criminal defamation laws and replace them with civil defamation laws (See, e.g., The

Special Rapporteur (E/CN.4/2000/63); The Special Rapporteur (E/CN.4/2001/64); The Special Rapporteur (A/67/357); Article 19, 2004) further stated that defamation should not be applied to the cases of criticism against public officials. The call aimed to avoid excessive restrictions on people's right to freedom and expression as defamation laws are often susceptible to abuse.

→ Regional Legal Instruments

● European Union (EU) Legal Instruments

The verbal pairing of 'illegal and harmful content' first appeared formally in a document produced by the EU in October 1996, namely the **Communication on Illegal and Harmful Content on the Internet** (Price, 2002). Through the document, the EU has shown their concern toward the importance of differentiating illegal and harmful content. They argued that the different content categories 'pose radically different issues of principle and call for very different legal and technological responses' (European Commission, 1996). The European Commission did not specify the definition of illegal content. However, several examples of illegal content mentioned include copyright infringement, libel, invasion of privacy, child pornography, dissemination of racist material, or incitement to racial hatred.

Furthermore, the European Commission was also aware that the exact definition of offenses for illegal content differs from country to country (European Commission, 1996). On the other hand, harmful content was defined as various types of material that may 'offend the values and feelings of other persons', including content expressing political opinions, religious beliefs, or views on racial matters. Nevertheless, the document further elaborated that what could constitute 'harmful' depends on cultural distinctiveness.

In the same year, the European Commission also issued a **Green Paper on the Protection of Minors and Human Dignity in Audio-visual**

and Information Services. Although it does not specifically distinguish 'illegal and harmful', the document offers a variant on the 'material banned for all by particular Member states' and 'certain material that might affect the physical and mental development of minors'. The first category primarily consists of child pornography, extreme gratuitous violence, and incitement to racial or other hatred, discrimination, and violence. On the other hand, the second category includes advertisements.

The **European Commission's Action Plan for the European Union for a Safer Use of the Internet 2007** states that illegal content is related to a wide variety of issues, such as instructions on bomb-making, which can threaten national security, child pornography, incitement to racial hatred, and libel. In comparison, harmful content is both that which is authorized but has restricted circulation (e.g., for adults only) and content that could be offensive to some users, even if publication is not restricted because of freedom of speech.

In mid-December 2020, the European Commission revealed the proposal for the Digital Services Act (DSA). The DSA proposal seeks to prepare an efficient regulation of innovative digital services in the internal market, promote online safety and the protection of fundamental rights, and establish effective governance in the supervision of intermediary services providers (European Commission, 2020). Passed by the EU in April 2022, the DSA can be used as a reference in discussing the distinction between illegal and harmful content.

In the explanatory memorandum, it is indicated that the multi-stakeholder consultation for the proposal resulted in a general agreement among the stakeholders, that: **'harmful content should not be defined in the Digital Services Act and should not be subject to removal obligations**, as this is a delicate area with severe implications for the protection of freedom of expression.' Thus, the DSA proposal

only defines 'illegal content' and does not contain a definition of 'harmful content'. Article 2 letter (g) of the DSA Proposal states the definition of 'illegal content' as follows:

'... 'illegal content' means any information, which, in itself or by its reference to an activity, including the sale of products or provision of services is not in compliance with Union law or the law of a Member State, irrespective of the precise subject matter or nature of that law.'

The absence of a clear line regarding the distinction between illegal and harmful content in the DSA proposal is appraised to make the definition of harmful content too broad and expand the discretion of intermediary service providers in its interpretation (Branden et al., 2021).

One study commissioned by the European Parliament titled 'Reform of the EU Liability Regime for Online Intermediaries' has also provided a reasonably sound foundation to distinguish between illegal and harmful content. Illegal content is defined as a large variety of information items that are not compliant with EU and national legislation, such as hate speech, incitement to violence, child abuse material, and revenge porn. On the other hand, harmful content is information that does not strictly fall under legal prohibitions, but might nevertheless have harmful effects, such as cyberbullying and mis/disinformation (Madiega, 2020).

● **Association of Southeast Asian Nations (ASEAN) Legal Instruments**

No single binding instrument regulates illegal and harmful content in ASEAN. Therefore, the distinction between illegal and harmful content also cannot be found. To date, ASEAN has only produced several types

of soft law, such as declarations, to regulate the forms of actions that can be classified as harmful content.

In 2017, ASEAN adopted the **ASEAN Declaration to Prevent and Combat Cybercrime**. The declaration contains the commitment of ASEAN member countries to collaborate in efforts to prevent and combat cybercrime. Furthermore, the declaration also acknowledged the importance of harmonizing laws related to cybercrime and electronic evidence at the regional level.

Furthermore, the 14th Conference of the ASEAN Ministers Responsible for Information (AMRI) in 2018 issued a **Framework and Joint Declaration to Minimize the Harmful Effects of Fake News**. The document provides several strategies that can be used by the state to combat fake news, starting from the creation of national laws, norms, and/or guidelines and increasing civil society involvement. However, the definition of fake news has also not been mentioned.

To conclude, several legal instruments explained above illustrate the importance of differentiating 'illegal' and 'harmful' content in regulating content on social media, considering the possible differences in handling and duty of care for illegal and harmful content in practice between states and platforms (Madiega, 2020). For example, stakeholders handle 'illegal content' by taking it down from social media platforms. Then, the possible suspect may be processed based on existing legal provisions. Moreover, 'harmful content' may be processed in accordance with the community guidelines of each social media platform (Vogelezang, 2020). In the regional context, an independent body can also be formed to overcome the disharmony of community guidelines between platforms and regulations between countries (De Streel, 2020). Further, the distinction between illegal and harmful content may also be advantageous in balancing internet users' rights while keeping online platforms accountable to regulators (Vogelezang, 2020).

International Norms and Standards on Regulating Illegal and Harmful Content

It is undeniable that freedom of expression is part of human rights, as reflected in Article 19 of the Universal Declaration of Human Rights (UDHR). Article 19 of the UDHR reads as follows:

'Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.'

Needless to say, the UDHR does not explicitly refer to freedom of expression on the Internet. It is unreasonable to think that the drafters had already imagined today's situation to that extent. However, the European Court of Human Rights refers to its interpretation of Article 19 of the UDHR 'in the light of present-day conditions'. The realm of the Internet has undoubtedly become a big part of today's 'conditions' of communicating (Benedek & Kettemann, 2013). Thus, freedom of expression on the Internet could also fall under the ambit of Article 19 of the UDHR.

The Internet and social media are a means for people to communicate with each other. In communicating, people certainly express various expressions in themselves. Unlike in the real world, people's expressions of communication on the Internet appear in the form of 'content'. Disclosure of expression in the realm of the Internet also needs to be governed. Therefore, there need to be regulations regarding illegal and harmful content on the Internet. The biggest challenge of this arrangement is how to formulate regulations to protect people against harmful content while ensuring freedom of expression. Thus, there needs to be a distinction between illegal and harmful content (Burns, 2020).

The IHRL addresses illegal and harmful content issues. In this case, IHRL has two functions. The first function is to mandate the prohibition of certain forms of expression, while the second function is to become a barrier to types of content that might be prohibited by the state (Sander, 2021). As mentioned above, the Universal Declaration on Human Rights Law (UDHR) is a key document that inspired numerous legal mechanisms addressing human rights, including the most relevant international treaties regulating human rights. The declaration has become a reference for states in formulating their national human rights regulations, including regulations relating to illegal and harmful content. The UDHR is also further elaborated in other international legal instruments such as the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), the International Covenant on Economic, Social and Cultural Rights (ICESCR), and the International Covenant on Civil and Political Rights (ICCPR) which Indonesia has also ratified through Law Number 12 of 2005 on Ratification of the ICCPR.

Several international conventions above could at least become the basis to regulate illegal and harmful content. Article 4 of ICERD requires states to criminalise all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any racial or ethnic groups. A similar provision can be found in Article 20(2) of ICCPR, which specifies that states are required to prohibit 'any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence'. Under that respective provision, the United Nations (UN) special rapporteur on freedom of expression (A/74/486) broadened the scope of hate speech beyond 'national, racial or religious hatred', extending it to adverse actions on the grounds of 'race, colour, sex, language, religion, political or other opinions, national or social origin, property, birth or other status, including indigenous origin or identity, disability, migrant or refugee status, sexual orientation, gender identity or intersex status'. The Rabat Plan of Action, adopted by a high-level group of human rights experts, suggests that states should conduct a six-part threshold test in applying Article 20(2) of the ICCPR. The six thresholds include context, speaker, intent, content and

form, the extent of the speech act, and likelihood. Figure 1, below, provides brief explanations of each threshold mentioned above.

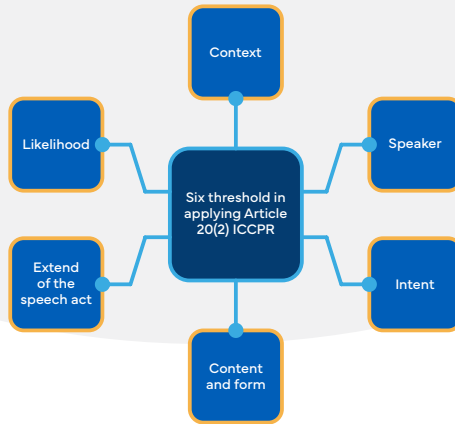


Figure 1. Six-Part Thresholds Test in Applying Article 20(2) of the ICCPR
(Source: The Rabat Plan of Action 2011)

- **Context:** The speech in question shall be analysed within the social and political context prevalent at the time the action was made and disseminated.
- **Speaker:** The actor who made or disseminated the speech shall be analysed for their position or status in society.
- **Intent:** Negligence and recklessness are not sufficient to fulfill the threshold. Therefore, the triangular relationship between the object and subject of the speech act as well as the audience shall be further analysed.
- **Content and form:** The content and form of the speech shall be analysed to be able to examine the intent and risk of harm of the speech.
- **Extent of the speech act:** The reach of the speech act shall be analysed, including the speech magnitude and size of its audience.
- **Likelihood:** Some degree of risk of harm arising from the speech must be identified.

Furthermore, Article 19 of ICCPR became the basis for the right to freedom of expression and the state's responsibility in protecting freedom of expression. The UN Human Rights Committee (HRC) also elaborated that those countries must adhere to the standards elaborated in Article 19(3) of the ICCPR, which states as follow:

'The exercise of the [right to freedom of expression] carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals.'

According to General Comment No. 34 on Article 19 of the ICCPR, the regulation related to freedom of expression in Article 19 of the ICCPR also includes all forms of audio-visual as well as electronic and internet-based modes of expression. However, the UN Human Rights Committee Joint Declaration on Freedom of Expression further noted that the currently available regulatory approaches in telecommunications could not be easily applied in the context of the Internet (Article 19, 2021).

Another report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/HRC/38/35) provides that states may not restrict the right to hold opinions without interference. In applying Article 19(3) of the ICCPR, the state limitations on freedom of expression must meet the following conditions:

Legality: the restrictions must be adopted by a regular legal process and limit government discretion in a manner that distinguishes between lawful and unlawful expression with 'sufficient precision'.

- **Necessity and proportionality:** states must demonstrate that the restriction imposes the least burden on the exercise of the right and actually protects, or is likely to protect, the legitimate State interest at issue.
- **Legitimacy:** any restriction must protect only those interests specified in Article 19(3), including the rights of reputations of others, national security or public order, or public health or morals.

In addition to Article 19(3) ICCPR, the cumulative conditions above should be satisfied by the state in applying Article 20(2) ICCPR, which requires states to prohibit '*advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence*'.

The report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/74/486) further describes several types of regulations that are often adopted by the state but do not meet the requirements stated in Article 19(3) of the ICCPR. One of them is anti-blasphemy laws.

In addition to the aforementioned binding legal instruments, the UN also issued the UN Strategy and Plan of Action on Hate Speech. Although its nature can be classified as soft law, the document provides a definition of hate speech that can be used as a basis for discussions about illegal and harmful content. The UN Strategy and Plan of Action on Hate Speech defines hate speech as any kind of communication in speech, writing, or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factors. The document further affirmed that hate speech includes as an act that is not prohibited but may be harmful.

Recently, the Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/HRC/47/25) provides the definition of disinformation that can be used

as an additional basis for discussions about illegal and harmful content. It is stated that disinformation is understood as false information that is disseminated intentionally to cause serious social harm, and misinformation as the dissemination of false information unknowingly. It is further explained that some forms of disinformation can amount to incitement to hatred, discrimination, and violence, which are prohibited under international law. Additionally, the document reaffirms that it is vital to clarify the concepts of disinformation and misinformation within the framework of IHRL.

The aforementioned report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/HRC/38/35) also produces several recommendations for countries in protecting the right of opinion and expression on the Internet, namely that: (1) states should repeal any law that criminalizes or unduly restricts expression; (2) states should only seek to restrict content pursuant to an order by an independent and impartial judicial authority and in accordance with due process and the cumulative conditions of legality, necessity, and legitimacy; (3) states should not adopt models of regulation where the government, rather than judicial authorities, become the arbiters of lawful expression. As for the information and communications technology (ICT) companies, the report recommends that they should recognize that the authoritative global standard for ensuring freedom of expression on their platforms is human rights law and not the varying laws of states of their own private interest.

Furthermore, the Budapest Convention on Cybercrime is the first international treaty that specifically addresses the issue of internet and computer crime. Although it does not distinguish illegal and harmful content, the Convention regulates several acts where the parties shall adopt such legislative and other measures as may be necessary to establish as criminal offenses under their domestic law. This act includes offenses related to child pornography and offenses related to infringements of copyright and related rights.

Based on several legal instruments listed above, it can be concluded that there are no international treaties that specifically regulate illegal and harmful content on the Internet. However, international legal instruments that could serve as a basis to regulate harmful content are actually available. Unfortunately, the arrangements are still scattered in several international treaties, even in some 'soft law' documents. At least a comprehensive international legal instrument regulating harmful content is needed to effectively respond to the growing use of social media.

Additionally, while not necessarily listed under international legal instruments, ensuring transparency in regulating content is deemed more desirable. Several studies argued that increasing transparency in content moderation is vital in ensuring more effective content monitoring and moderating practices (WEF, 2021; Gorwa et al., 2020; APC, 2018; De Gregorio, 2020). Additionally, the United Nations' Special Rapporteur emphasized the importance of upholding international human rights frameworks in moderating content online (United Nations Human Rights Office of the High Commissioners [OHCHR], 2018). In this aspect, the EU model of content moderation, through the EU Digital Service Act (DSA), also emphasizes fundamental rights and proportionality principles in regulating content within its jurisdiction (European Parliament, 2020).

Unfortunately, there is an apparent gap between the best practice and action states take as they resort to a more punitive approach. This approach often neglects the human rights principles and, as a consequence, can be used to limit the freedom of speech and might potentially be used as a tool for censorship (Banchik, 2020). For instance, Germany's NetzDg compels content removals based on its criminal code provisions, raising severe freedom of expression concerns with its broad concept of 'defamation of religion' and 'hate speech' provisions (Article 19, 2017). Unfortunately, several other states, such as India, Kenya, Malaysia, the Philippines, Russia, Turkey, and Venezuela, enact similar content regulations that adopt a punitive mechanism approach in their way to regulate digital content (WEF, 2021). This shows how states actively choose to apply the punitive

mechanism approach—an approach that is distinct—even contradict the proposed norms on content regulations. Hence highlighting the stark and fundamental gap between the proposed norms on content regulations, as found in several studies and suggestions from the rapporteur, and the actual practices implemented by states, as shown in the wide adoption of punitive content regulation.

Platforms and Content Moderation

Social media platforms (e.g., Facebook, Instagram, Twitter, and YouTube) possess and benefit financially from the vast amounts of users and the data they produce (Maulana, 2021). Furthermore, they are also the primary governor of user and community practice within their platform. As they hold a vast amount of power, the public should hold them to some form of responsibility to protect them (Doctorow, 2021b), especially if the benign stated goals of these corporations are to be believed.

Nevertheless, recent developments and extant literature would suggest that these corporations would go out of their way to ensure profitability rather than the interests of their users (e.g., Aschoff, 2020; Marcetic, 2021; Oates, 2020). The idea of growth, engagement, and scale are all paramount to social media platforms to attract advertisers and investors. 'Daily Active Users' and 'Monthly Active Users' are all essential indices of success for social media companies, often above other values such as social development, safety, quality, and meaningful interaction (Zulli et al., 2020). Data extraction, behavioural modification, and externalization of costs are all dully unsurprising as tech firms, like any other firms, are driven by the need to assure long-term profitability (Morozov, 2019).

The profit-oriented approach may negatively affect their content moderation practices as platforms try to find ways to increase users' engagement and screen time on their site. Reviglio and Astoti (2020, p. 4) found that the social and political polarization often found on social media is actually part of their business model. They do so by creating a content

feed that is excessively homophily, resulting in polarized clusters and divisive content. As these clusters are emotionally and politically charged, they tend to defend their beliefs and attack others. One example is during the 2017 Jakarta Gubernatorial Election when supporters of the two candidates—the incumbent Vice Governor turned Governor Basuki Tjahja Purnama (Ahok), and Anies Baswedan—participated in what Merylina Lim (2017) called tribal nationalism. Driven by racial and religious signifiers, the supporters of both candidates extrapolated their divisiveness from offline political activity to the online world, creating algorithmic enclaves.

Lim (2017) posits algorithmic enclaves as techno-socially constructed imagined communities built when groups of individuals, facilitated by their constant interaction with algorithms, try to form a (perceived) sense of identity online, where they would defend their beliefs and attack others. Even though algorithms themselves do not predetermine the creation of these enclaves, rather, they are mutually shaped by user-algorithm interactions. Furthermore, Lim (2017, p. 13) contends that it is visible to see these enclaves online in the fragmented social space between 'Chinese-Christians, hijabis (both pro and anti-Ahok), and pribumi (native).'

Another criticism of platforms' self-regulatory mechanisms is the lack of transparency in their method of regulating content. A considerable amount of content generated every minute pushes platforms to use advanced monitoring tools, such as algorithms and AI machines, to monitor and regulate their content. Algorithms are often used to flag content that violates community standards. Left to their own devices, algorithms can either flag and remove too much content or leave negative speech intact because they are not sophisticated enough to parse through minority languages and regional dialects (Sinpeng & Martin, 2021).

Gorwa et al. (2020) also observed that automated moderations using advanced monitoring tools by the platforms often result in difficulties in auditing the process and complicated issues of justice (as some views and values from certain groups are privileged) depoliticize the political process

in the platform content moderation. They argue that what matters is not how to improve platforms' 'efficiencies in tracking and to take down harmful content but the process of how they determine something as harmful.

Moreover, the Association for Progressive Communications (APC, 2018) found that many platforms enter undisclosed and non-transparent agreements with states in moderating and even removing content. This non-transparent agreement is argued to bypass democratic institutions and facilitate censorship of legitimate speeches. One of the prominent examples is cooperation between Israel and Facebook to tackle 'incitement' content where the term is vaguely defined, whereas one of the interpretations suggests that it is content that resists and criticizes Israel's policies (APC, 2018).

There are also labour issues in the platform's content moderation practices. Aside from using advanced monitoring tools, platforms also enact manual, labour-based practices to monitor content on their sites. Facebook employs 15,000 workers, Google employs around 10,000 workers to oversee their range of products—such as YouTube—and Twitter has approximately 1,500 moderators. Unfortunately, the vast majority of these content moderators are outsourced workers, resulting in several issues, such as underpaid labour and gruesome—even traumatic—working conditions (Barrett, 2020).



CHAPTER I

Regulations of Online Harmful Content in Indonesia



The first chapter will map and analyse Indonesian regulations related to illegal and harmful content. First, the constitutional basis associated with illegal and harmful content regulations, especially those related to constitutional rights and limitations, are examined. Second, the author analyses illegal and harmful content classification based on existing regulations, specifically defamation, hate speech, and false news provisions. Third, applicable remedies and mechanisms for illegal and harmful content handling are explained. Finally, the responsibilities of social media platforms based on the current regulations are highlighted.

Constitutional Basis on Regulating 'Illegal and Harmful Content' in Indonesia

As a state based on the rule of law, Indonesian regulations play an essential role in regulating various aspects of people's lives, including how they interact with others. Hence, state regulations become the primary basis for identifying what constitutes illegal and harmful content and how to handle it.

Furthermore, the discussion on illegal and harmful content is inseparable from the context of human rights protection. The regulation on these contents might potentially limit the enjoyment of citizens' human rights, especially the freedom of expression. Regarding harmful content regulations, the 1945 Constitution of the Republic of Indonesia (the 1945 Constitution) has explicitly guaranteed the protection of human rights, including freedom of expression and freedom of information, as follows:

Article 28

The freedom to associate and assemble, express written and oral opinions, etc., shall be regulated by Act.

Article 28E

(3) *Every person shall have the right to the freedom to associate, assemble, and express opinions.*

Article 28F

Every person shall have the right to communicate and obtain information for the purpose of developing his/herself and social environment and shall have the right to seek, obtain, possess, store, process, and convey information by employing all available types of channels.

These provisions, especially Article 28E and Article 28F included in the second amendment of the 1945 Constitution, are adopted from Articles 19, 20, and 21 of the Universal Declaration of Human Rights (UDHR), which contain similar wording. It was confirmed by the constitution drafters in the 2nd amendment of the 1945 Constitution in 2000 (Constitutional Court, 2010).

To ensure the implementation, Article 28I of the 1945 Constitution also provides constitutional obligations to the State, especially the government, to protect, promote, enforce, and fulfill human rights. It is generally understood that the State has the primary responsibility of protecting human rights. Thus, the State should provide protection and guarantee to realize the citizens' human rights (Nickel, 1993). The State's role was also emphasized, for instance, in the International Covenant on Civil and Political Rights, which provides the general responsibility of the state for protecting human rights. The protection of human rights provided by the State can be accomplished through the political process and the formulation of legal instruments (Evans, 2009).

Furthermore, the legal instruments formulated also act as a lawful means to limit the enjoyment of human rights. In Indonesia, the 1945 Constitution provided that the enjoyment of human rights could only be limited through Acts. The details are stated in Article 28I of the 1945 Constitution as follow:

*'In the exercise of their rights and freedom, every person shall abide by the **limitations provided by the Acts** to solely guarantee the recognition and respect for the rights and freedoms of the others and to **comply with just demands in accordance with considerations for morality, religious values, security, and public order in a democratic society.**'*

The Indonesian Constitution's take on the limitation of human rights is also inspired by the UDHR,⁵ which further elaborated in the International Covenant on Civil and Political Rights (ICCPR).⁶ Hence, it is crucial to ensure that the legal instruments formulated by the state could provide robust legal protection and guarantee human rights, including the freedom of speech and expressions that are often 'attacked' in this digital age.

The analysis of the Acts will also elaborate on whether the definition or explanation of the concerning illegal and harmful content is clear enough and whether the related Acts comply with the constitutional requirements, including the concern on morality, religious values, security, and public order. Furthermore, how the regulations handle these contents will also be analyzed. Additionally, various court decisions related to the concerned Acts will be explored to understand further how courts interpret illegal and harmful content provisions. It is essential to fully understand how courts implement the law in deciding cases related to these contents.

⁵ It is worth noting that the Universal Declaration of Human Rights inspired the formulation of human rights provisions in the 1945 Constitution. It is elaborated in the Naskah Komprehensif Perubahan UUD 1945 Buku VIII. See Article 29 paragraph 2 of the Universal Declaration of Human Rights.

⁶ See e.g., Article 11 paragraph 3, Article 18 paragraph 3, Article 19 paragraph 3 letter (b), and Article 22 paragraph 3). Other regional conventions also have similar provisions, such as the European Convention on Human Rights (See Article 8 paragraph 2, Article 9 paragraph 2, and Article 10 paragraph 2), the American Convention on Human Rights (See Article 12 paragraph 3, Article 13 paragraph 2 letter b, and Article 15). Furthermore, Indonesia has also ratified the ICCPR through Law No. 12 of 2005.

Classification of Illegal and Harmful Content in Indonesian Regulations

In Indonesia, the EIT Act could be said to be the primary legal basis for regulating online content. This Act essentially regulates myriad aspects of cyberspace, including digital content. This Act is also widely used for various crimes related to cyberspace, specifically those associated with illegal and harmful content. Although the EIT Act did not explicitly define 'content', it could be interpreted as electronic information and electronic documents according to the EIT Act. It describes both as follows:⁷

Electronic Information** means **one cluster or clusters of electronic data**, including but not limited to writings, sounds, images, maps, drafts, photographs, electronic data interchange (EDI), electronic mail, telegrams, telex, telecopy or the like, letters, signs, figures, Access Codes, symbols or perforations that **have been processed for meaning or understandable to persons qualified to understand them.

Electronic Documents** means **Electronic Information that is created, forwarded, sent, received, or stored** in analogue, digital, electromagnetic, optical form, or the like, visible, displayable, and/or audible via Computers or Electronic Systems, including but not limited to writings, sounds, images, maps, drafts, photographs or the like, letters, signs, figures, Access Codes, symbols, or perforations **having particular meaning or definition or understandable to persons qualified to understand them.

⁷Article 1 number 1 and Article 1 number 4 of Law No. 11 of 2008 jo. Law No. 19 of 2016 on Electronic Information and Transaction.

Similar wording could be found in sectoral Acts, such as the Public Information Disclosure Act (Law No. 14 of 2008), which defines information as 'statement, ideas, and signs having a value, meaning and message, both the data, fact, and clarification that can be seen, heard and read, and are presented in various packages and formats, in accordance with the development of the information and communication technology, both electronically and non-electronically'.⁸ Therefore, any electronic information and electronic documents containing potentially illegal and harmful substances could be categorized as 'illegal' and 'harmful' content.

Essentially, the EIT Act as the primary legal basis in regulating content does not differentiate between illegal and harmful content. However, the General Elucidation of the 2016 EIT Act uses the term 'illegal content' (*konten ilegal*) to refer to electronic information and/or electronic documents containing:⁹

- content that violates decency;
- gambling content;
- slander or defamation;
- extortion and/or threats;
- false and misleading news resulting in consumer losses in electronic transactions;
- hatred or hostility based on ethnicity, religion, race, and class; and
- threats of violence or intimidation that are directed to an individual.

⁸Article 1 number 1 Law No. 14 of 2008 on Public Information Disclosure.

⁹See General Elucidation of Law No. 19 of 2016 on the Amendment of Law No. 11 of 2008 on Electronic Information and Transaction.

These illegal content are substantially regulated in Chapter VII regarding Prohibited Acts (*Perbuatan yang Dilarang*) of the EIT Act, specifically from Article 27 to Article 29. Other Articles in this chapter are provisions concerning *Title 1 and Title 2 Cybercrime*¹⁰ such as hacking, cracking, and phishing.¹¹

Thus, although several international and regional legal instruments mentioned above highlighted the importance of differentiating between 'illegal' and 'harmful', in the EIT Act context, potentially harmful content mentioned in several international or regional legal instruments above are essentially classified as illegal content according to the EIT Act. Consequently, as the prohibited acts are considered a criminal offense, the offender can be criminally prosecuted.

Although the EIT Act has become the primary legal basis for regulating online content, it cannot stand alone. According to the Constitutional Court of the Republic of Indonesia, the EIT Act essentially expands the scope of offense from physical space to cyberspace.¹² Therefore, other Acts should be referred to determine whether online content is deemed illegal/harmful in some instances.

The scope of illegal content mentioned above is further expanded in the Government Regulation on Electronic System and Transaction Implementation (GR ESTI) as the primary implementing regulation of the EIT Act.¹³ The GR ESTI implicitly 'differentiates' types of content into two categories, namely: (a) content that violates laws and regulations; and (b) content that disturbs community and public order.

The first classification covers content 'that are prohibited in accordance with the provisions of laws and regulations' (*memuat muatan yang dilarang sesuai dengan ketentuan peraturan perundang-undangan*),

¹⁰See Article 2 to Article 8 of the Convention on Cybercrime, Budapest, 23.XI.2001.

¹¹See Article 30 to Article 37 of Law No. 11 of 2008 jo. Law No. 19 of 2016 on Electronic Information and Transaction.

¹²See e.g., Constitutional Court Decision No. 50/PUU-VI/2008 concerning Review of Law No. 11 of 2008 on Electronic Information and Transaction against the 1945 Constitution of the Republic of Indonesia.

¹³See Article 95 and Article 96 of Government Regulation No. 71 of 2019 on Electronic System and Transaction Implementation

which essentially means 'illegal content'. The first category includes electronic information and electronic documents containing elements of (a) pornography, (b) gambling, (c) slander, (d) defamation, (e) fraud, (f) hatred towards community, based on ethnic groups, religions, races, and inter-groups, (g) violence and/or violence against children, (h) violations of intellectual property, (i) violations of trade in goods and services through electronic systems, (j) terrorism and/or radicalism, separatism and/or prohibited dangerous organizations, (k) breaches of information security, (l) violations of consumer protection, (m) breaches in the health sector, (n) violations of supervision of medicine and food.¹⁴

Most of the content types mentioned above are substantially regulated as prohibited acts in the EIT Act, including pornography, gambling, slander, defamation, fraud, and hatred towards the community, based on ethnic groups, religions, races, and inter-groups. Even online content containing violence and/or violence against children, terrorism and/or radicalism, separatism, and/or prohibited dangerous organizations could constitute prohibited acts in some instances. Other types of content should also refer to other Acts, such as the Criminal Code, the Copyright Act, the Health Act, the Food Act, and the Consumer Protection Act.

The second category regarding content that disturbs community and public order could be interpreted as 'harmful', as it seems that it refers to content that might not violate laws and regulations but could still be harmful to the community and public order. However, the formulation of this provision is deemed too broad, and even the explanation is unclear (this will be further elaborated on below).

Although the current legal framework mentions various types of content that can be classified as 'illegal' or 'harmful', this research will only examine the provisions relating to defamation, hate speech, and false news. Apart from the scope of the SM4P overall project, the selection of the content was also based on several reports from civil society organizations in

¹⁴See Article 96 letter a of the Government Regulation No. 71 of 2019 on Electronic System and Transaction Implementation.

Indonesia, which showcased that provisions related to defamation, hate speech, and false news are among the most used legal grounds for criminal prosecutions of someone's speech in social media (See, e.g., Tempo, 2020; and SAFEnet, 2021). Furthermore, this research also highlights one more classification of content that criticised and potentially limits freedom of expression in social media, namely content that disturbs community and public order. The elaboration is provided below.

➔ Slander and/or Defamation

The provision concerning slander and defamation in the EIT Act is considered one of the most used provisions (See e.g., Tempo, 2020; and SAFEnet, 2021). This provision is also the first EIT Act provision to be submitted for constitutional review.¹⁵ Furthermore, slander and defamation in the EIT Act are also inseparable from slander and defamation regulated in the Criminal Code.¹⁶ The Criminal Code defines the crime of slander and/or defamation as: '[...] deliberately attack[ing] someone's honour or reputation by accusing someone of something, with the obvious intent to give publicity.'¹⁷

This interpretation is affirmed by the Constitutional Court in Case No. 50/PUU-VI/2008, in which the Court states that the interpretation of defamation or slander in Article 27 paragraph (3) of EIT Act is inseparable from criminal law norms enshrined in Article 310 and Article 311 of the Indonesian Criminal Code.¹⁸ The Court held that:¹⁹

¹⁵See Article 281 to Article 2956 of the Indonesian Criminal Code.

¹⁶See Article 300, Article 303 and Article 303 bis of the Indonesian Criminal Code.

¹⁷See e.g., District Court of Mataram Decision No. 265/Pid.Sus/2017/PN.Mtr (12 June 2017) 29.

¹⁸See e.g., District Court of Semarang Decision No. 652/Pid.Sus/2020/PN Smg (23 December 2020) 42.

¹⁹Ibid.

*[3.17.1] That apart from the Court's considerations as described in the previous paragraph, the validity and interpretation of **Article 27 paragraph (3) of the EIT Act is inseparable from the primary legal norms in Article 310 and Article 311 of the Criminal Code, as the genus of delict** which requires a complaint (klacht) to be prosecuted, must be treated against prohibited acts in Article 27 paragraph (3) of the EIT Act, and therefore the a quo Article must also be interpreted as an offense requiring a complaint (klacht) to be prosecuted before the court*

Moreover, neither EIT Act nor the Criminal Code provides a sufficient explanation of what constitutes an Act's element that has 'insulting and/or defamatory' content. For instance, in case No.132/PID.B/2010/PN.MRK, the judge merely interpreted the element of 'deliberately attacking someone's honour or reputation by accusing someone of something' as an intention of the defendant to attack someone's honour or reputation by accusing someone of something.²⁰ The Constitutional Court, in the decision mentioned above, also reaffirmed that there must be a distinction between 'intentionally committing an act' and 'intentionally attacking the honour or good name of another person' in Article 310.²¹

Furthermore, in the absence of a clear definition or interpretation of defamatory content, several courts broaden the object of defamation, including the honour of a legal entity or state institution.²² For instance, in Case No. 223/Pid.Sus/2018/PN Kbm, the Court refers to the Supreme Court Decision No. 183 K/Pid/2010, which states that a legal entity could be an object of defamation.²³

²⁰Ibid.

²¹See District Court of Wonosari Decision No. 89/Pid.Sus/2020/PN.Wno (2 November 2020) 19.

²²See e.g., District Court of Kebumen Decision No. 223/Pid.Sus/2018/PN Kbm (17 December 2018) 41.

²³See Supreme Court Decision No. 183 K/Pid/2010 (20 May 2010) 15.

With various polemics that occurred due to the different interpretations of the police, prosecutor, and judiciary towards prohibited acts in the EIT Act, the Minister of Communications and Informatics (MOCI), the Attorney General, and the Chief of the Indonesian National Police issued guidelines for implementing the EIT Act (Joint Decree).²⁴ Specific for Article 27 paragraph (3), the Joint Decree emphasized that the focus of sentencing for this Article is not on the victim's feelings but the perpetrator's actions.²⁵ Furthermore, this Joint Decree states that the victim as a whistle-blower must be an individual with a specific identity and not an institution, corporation, profession, or position.²⁶ At a glance, this seems to contradict the Supreme Court Decision No. 183 K/Pid/2010 mentioned above. Nevertheless, the Joint Decree did not specify whether an institution could be addressed as an object of defamation or slander. Thus, it can still be interpreted that an institution could be an object of defamation or slander. However, the whistle-blower should be a specific individual. The formulation of slander/defamation content in the EIT Act has affected the implementation of freedom of expression in Indonesia and has been highly criticised by various civil society organisations.²⁷

→ False and Misleading News (Misinformation/Disinformation)

● General False and Misleading News (Criminal Code)

Misinformation/disinformation, in general, is regulated in the Criminal Code. Articles that govern the prohibition of spreading false news or information are, among others, Article 14 paragraph (1) and (2), and Article 15 of the Criminal Code. However, those Articles only stipulate the term as 'false news or information which can cause trouble among the people'. These Articles did not further explain what types of news or information constitutes 'false news or information and what degree of

²⁴Joint Decree of the Minister of Communications and Informatics, the Attorney General, and the Chief of the Indonesian National Police No. 229 of 2021, No. 154 of 2021, No. KB/2/VI/2021.

²⁵Ibid, 12.

²⁶Ibid.

²⁷Further explanation of the concerns, trends, and impact of defamation law will be provided in Chapter II and Chapter III.

trouble can cause problems among the people' referred to in the articles. In the absence of such stipulations, it is essential for us to take a look at the jurisprudence concerning these articles.

Because of its broad nature compared to misinformation/disinformation articles in the EIT Act, these articles are also used for cyberspace cases, mainly related to misinformation. For example, in Case No. 203/Pid.Sus/2019/PN.Jkt.Sel regarding the case of misinformation in social media, the prosecutor used alternative charges using Article 14(1) of the Indonesian Criminal Code and Article 28(2) of the EIT Act regarding hate speech.

Furthermore, the Court held that the defendant was proven to have committed the crime as referred to in the first indictment. The Court, however, did not explicitly define what is 'false news' but associated it with the facts presented by the witnesses. The Court was more concerned about the impact that occurred because the 'news' went viral and became a trending topic. According to the Court, virality became the 'seeds' for the chaos that has emerged to the surface.²⁸

Meanwhile, Case No. 471/Pid.Sus/2020/PN.Bdg defines 'false news' as information that is fabricated or conveyed untruthfully—not in accordance with facts and intentionally made to mislead the information recipient.²⁹ The Court emphasized that the intention of the publication of 'false news' is crucial, as 'false news' should be published to make people believe the content is genuine.³⁰ Several decisions above show that there are different interpretations of the terms contained in Articles 14(1), 14(2), and 15. The absence of such clear definitions and measures can then lead to the arbitrary application of rules by the authorities.

²⁸ See District Court of South Jakarta Decision No. 203/Pid.Sus/2019/PN.Jkt.Sel, p.142, para. 7.

²⁹ See District Court of Bandung Decision No. 471/Pid.Sus/2020/PN.Bdg, p.121, Ad. 2.

³⁰ *Ibid.*, p.130, Ad. 3

Furthermore, according to GR ESTI, misinformation/disinformation content is also classified as content that disturbs community and public order.³¹ Although it is classified under different categories with content that violates laws and regulations, based on the explanation above, content on disinformation is still classified as content that violates the Criminal Code. In some instances, it could also be used in parallel with provisions concerning prohibited acts in the EIT Act.

False and Misleading News Resulting in Consumer Losses in Electronic Transactions (the EIT Act)

False and misleading news in Article 28 paragraph (1) of the EIT Act is different from misinformation/disinformation in the Criminal Code.³² The elements provided in the EIT Act are more narrowly defined compared to the Criminal Code, as it only covers false and misleading information in the electronic transaction context. Article 28 paragraph (1) of the EIT Act is intended to protect consumers in electronic transactions. According to the EIT Act, the definition of an electronic transaction is 'a legal act that is committed by the use of computers, computer networks, and/or other electronic media.'³³ Therefore, a person merely distributing or transmitting disinformation or fake news is not sufficient to be classified as a prohibited act under this article. The element of 'resulting in consumer losses' should also be proved.³⁴

The content under this Article could also fall under the category of content that contains violations of trade in goods and services through electronic systems. Furthermore, the definition of 'consumer' in the context of Article 28 paragraph (1) of the EIT Act should refer to Law No. 8 of 1999 on Consumer Protection (Consumer Protection Act). The Act defines a consumer as 'users of goods and/or services available in the

³¹See Article 96 letter b and Elucidation of Article 96 letter b of Government Regulation No. 71 of 2019 on Electronic System and Transaction Implementation.

³²Ibid, 16.

³³See Article 1 number 3 of Law No. 11 of 2008 on Electronic Information and Transaction.

³⁴See e.g., District Court of East Jakarta Decision No. 532/Pid.Sus/2020/PN.Jkt.Tim (6 October 2020) 21–22; District Court of Bale Bandung Decision No. 84/Pid.Sus/2021/PN Blb (19 April 2021) 17–19.

community, both for the benefit of themselves, their families, other people, and other living creatures and not for trading.³⁵ Moreover, false and misleading news in the EIT Act could also be related to Article 378 of the Criminal Code that regulates fraud.

→ **Hatred or Hostility Based on Ethnicity, Religion, Race, and Intergroup (Hate Speech)**

Provision concerning hate speech in Article 28 paragraph (2) of the EIT Act is also one of the most used Articles in the EIT Act (See e.g., Tempo, 2020; and SAFEnet, 2021). The Article concerning hate speech could be associated with several Acts, including the Criminal Code, the Elimination of Race and Ethnic Discrimination Act, and the Prevention of Religious Abuse and/or Blasphemy Act.

Under the Indonesian Criminal Code, some articles govern the prohibition of the expression of feelings of hostility, hatred, or insults, namely Article 156, Article 156a, and Article 157. Article 156 of the Indonesian Criminal Code is actually the 'parent' of Article 156a. Article 156 of the Indonesian Criminal Code regulates acts against 'something or several groups of residents of the State of Indonesia'. Meanwhile, Article 156a regulates actions against one of these groups, namely the religious group (Indonesian Institute of the Independent Judiciary, 2018).

The related terminologies could refer to the Elimination of Race and Ethnic Discrimination Act. This Act provides many forms that could be constituted as 'race and ethnic discrimination', including: making writings or pictures to be placed, pasted, or distributed in public places or other places that can be seen or read by others; giving a speech; expressing, or saying certain words in a public place or other places that can be heard by others; wearing something in the form of objects, words, or pictures in public places or other places that others can read; or committing the deprivation of people's lives, such as torture, rape, obscene acts, theft with violence, or deprivation of liberty based on racial and ethnic discrimination.

³⁵See District Court of North Jakarta Decision No.1537/Pid.B/2016/PN.JKT.UTR, p. 594.

Several discourses concerning hate speech Articles include the terms 'in public', 'with the intention that the contents are known or more publicly known', and 'intergroup'. For instance, in the context of 'in public' terminology, the judge in case No. 1537/Pid.B/2016/PN JKT.UTR interpreted 'in public', referring to the book 'Delik - Delik Khusus Kejahatan Terhadap Kepentingan Hukum' by Drs. PAF Lamintang, S.H., which states that 'in public' in the criminal formulation regulated in Article 156a of the Indonesian Criminal Code does not mean that the feelings expressed by the perpetrator or the actions committed by the perpetrator must always occur in a public place, but it is sufficient if the emotions expressed by the perpetrator can be heard by the public, or actions that are carried out by the perpetrator can be seen by the public. The Elucidation of the Prevention of Religious Abuse and/or Blasphemy Act also refers the term 'in public' to the terminology used in the Criminal Code.

Regarding the second term, the intention to make content more publicly known can be derived through an inner attitude towards an awareness that their actions can indeed be known by the public (Purwati, 2018). A person's inner attitude is certainly not something that is easy to prove. Therefore, the intention is considered to be reflected when the perpetrator performs their actions in public. Basically, the intent here is to emphasize the phrase 'in public' contained in Article 157.

Regarding the term 'intergroup', although it could refer to several Acts mentioned above, in practice, this term could be broadly interpreted by the police, prosecutor, and judiciary. The Constitutional Court also acknowledges the broad interpretation of 'intergroup'. In Case No. 76/PUU-XV/2017, the Court held that:³⁷

³⁷Constitutional Court Decision No. 76/PUU-XV/2017 concerning Review of Law No. 11 of 2008 on Electronic Information and Transaction as Amended by Law No. 19 of 2016 on the Amendment of Law No. 11 of 2008 on Electronic Information and Transaction (27 March 2018) 66–68

[3.13.1] Whereas the term 'intergroup', according to the Court, **is not clear and firm**. The meaning of the **term cannot be immediately known**, in contrast to the terms 'ethnic', 'religion', and 'race', which are placed in parallel with the term 'intergroup' and even give rise to a popular abbreviation in society, namely SARA. Although it is unclear and firm, it does not mean that the 'intergroup' does not exist. [...]

The term 'intergroup' clearly does not refer to 'intergroup' as referred to in Article 163 and Article 131 IS, but rather to the sociological reality of the existence of 'other groups' outside of ethnicity, religion, and race. [...]

Moreover, it is also complicated to distinguish which content is offending and intended to cause hostility and hatred. In Case No. 366/Pid.Sus/2019/PN.JKT.SEL, for example, the District Court of South Jakarta's decision held that offending content could be constituted as causing hatred and hostility.³⁸ The details are stated as follows:

[...] Considering, that based on the considerations, what is proven in this case is intentional with a conscious certainty, as in the present case, the Defendant understands or realizes that his actions, consequences, and the accompanying circumstances, that the Defendant's post will **offend Chinese ethnic and groups of Presidential and Vice-Presidential Candidate No. 1 supporters, and the President and Vice President Candidates No. 1 Joko Widodo and Ma'aruf Amin, as well as the KPU as the organizer of the 2019 general election.**

Moreover, it also illustrates that the Court broadly interprets the term 'group', which also includes groups of people (Presidential and Vice-

³⁸See e.g., District Court of South Jakarta Decision No. 366/Pid.Sus/2019/PN.JKT.SEL (15 August 2019) 84–85

Presidential Candidates) and state institutions (KPU). A broader interpretation of 'group' could be found in Case No. 16/Pid.Sus/2020/PN Jap, in which the District Court of Jayapura held that the Nation of Indonesia (Bangsa Indonesia) also constituted a 'group'.³⁹ The Court stated that the Defendant's post leads to hostility; in this case, hostility toward the nation of Indonesia.

In another instance, the District Court of Kendari held that a harsh critique and negative comment toward State institutions could also fall under Article 28 paragraph (2) of the EIT Act. The Court held that if a person understands that there will be various responses, both pro and contra, which cause social polemic through social media, the content they post could be constituted as content that causes hatred or hostility among groups of people.⁴⁰

This Article could also be used to target the publication or dissemination of content containing terrorism and/or radicalism, separatism, and/or prohibited dangerous organizations.

The absence of clear limitations on how content could be considered hatred or hostile has the high potential to violate citizens' freedom of expression protected by the 1945 Constitution. It could be seen from various interpretations by several district courts, as mentioned above, which could broaden the meaning of the articles.

Furthermore, in the context of publication through the press, the Press Act can be used as one of the legal bases, as it prohibits ads that result in degrading a religion and/or disrupting harmony between religious life, contrary to the sense of public decency. However, there is no explicit definition in the Press Act concerning what constitutes 'degrading a religion and/or disrupting harmony between religious life, and contrary to the sense of public decency'. Thus, the interpretation of those terms is left wide open and prone to be assessed subjectively.

³⁹See District Court of Jayapura Decision No. 16/Pid.Sus/2020/PN Jap (29 April 2020) 38–39.

⁴⁰See District Court of Kendari Decision No. 426/Pid.Sus/2021/PN Kdi (16 September 2021) 24. See also Joint Decree of Minister of Communications and Informatics, the Attorney General, and the Chief of the Indonesian National Police No. 229 of 2021, No. 154 of 2021, No. KB/2/VI/2021, 20–21.

➔ Contents that Disturb Community and Public Order

Although it seems that this content is different from content that violates laws and regulations according to Article 96 letter a of the GRE STI, this type of content could still be broadly interpreted. To date, no clear explanation of what 'disturbing community and public order' is provided at the Act level. Consequently, it could be broadly interpreted by state institutions and law enforcers. In several regional regulations, public order (*ketertiban umum*) is defined as 'a condition in which the government and the citizens could carry out their activities in an orderly manner'.⁴¹ In the context of Regional Regulation concerning public order, the scope of public order could vary depending on the region.

The elucidation of Article 96 letter b explains that what it means by 'disturbing the community and public order' includes, among others, falsified information and/or facts.⁴² Although publication of falsified information is not specified as a prohibited act under the EIT Act, it still can be associated with other prohibited acts under the EIT Act, such as hatred and/or hostility toward community, based on ethnic groups, religions, races, and inter-groups,⁴³ and it also falls within the scope of Article 14 of the Indonesian Criminal Code. Thus, although falsified information is not classified as content that violates laws and regulations under GRE STI, it still could be prosecuted based on the Criminal Code and the EIT Act. Therefore, the example mentioned in the explanation of Article 96 letter b is essentially constituted as content that also 'violates laws and regulations'.

Considering that Article 96 letter a and letter b is principally the same, differentiation of classification of content that constitutes 'violates laws and regulations' and 'disturbing the community and public order' becomes questionable. Further questions will arise regarding other types of content

⁴¹See e.g., Regional Regulation of Special Capital Region of Jakarta No. 8 of 2007 on Public Order; Regional Regulation of Special Region of Yogyakarta No. 2 of 2017 on Peace, Public Order and Community Protection; and Regional Regulation of Buton Regency No. 2 of 2020 on the Implementation of Peace, Peace, Public Order and Community Protection.

⁴²See Article 96 letter b of the Government Regulation No. 71 of 2019 on Electronic System and Transaction Implementation.

⁴³See e.g., District Court of South Jakarta Decision No. 366/Pid.Sus/2019/PN.JKT.SEL (15 August 2019) 77–78.

that disturb community and public order, as the explanation only mentioned one type of content. The terminology used is too vague and could be broadly interpreted in accordance with the government or state's interests. The elucidation only provides one example of what could be constituted as 'disturbing community and public order'. Therefore, the potential for human rights violations based on this provision is high, especially concerning freedom of expression in social media. If a piece of content is deemed disturbing community and public order, it could be banned by social media platforms based on the request from the government.

→ **Consequences of Online Content Regulations**

As aforementioned, according to the EIT Act, 'online content' can refer to any electronic information and documents. However, the police will also need to refer other laws and regulations in handling illegal and harmful content online. The table below summarizes the types of online content regulated in multiple Acts, including four categories of content explained above.

Table 1. Online Content Classification Based on Indonesian Regulations

Types of Content	Classification	Related Regulations ⁴⁴
Contents against propriety (including pornography and child pornography)	Crime	<ul style="list-style-type: none"> ● EIT Act 2008 jo. 2016 ● Criminal Code 1946 ● Pornography Act 2008 ● GR ESTI 2019
Contents of gambling	Crime	<ul style="list-style-type: none"> ● EIT Act 2008 jo. 2016 ● Criminal Code 1946 ● GR ESTI 2019

⁴⁴The regulations listed start from the primary/most relevant legal basis based on hierarchy of regulations, and continue with several related regulations. Apart from GR ESTI, other implementing regulations of specific Acts are not explicitly listed but can still be used as the basis for the technical implementation of those Acts.

Contents of slander and/or defamation	Crime	<ul style="list-style-type: none"> • EIT Act 2008 <i>jo.</i> 2016 • Criminal Code 1946 • GR ESTI 2019
Contents of extortion and/or threats	Crime	<ul style="list-style-type: none"> • EIT Act 2008 <i>jo.</i> 2016 • Criminal Code 1946 • GR ESTI 2019
False and misleading information resulting in consumer loss in Electronic Transaction	Crime	<ul style="list-style-type: none"> • EIT Act 2008 <i>jo.</i> 2016 • Consumer Protection Act 1999 • GR ESTI 2019
Information aimed at inflicting hatred or dissension on individuals and/or certain groups of community-based on ethnic groups, religions, races, and inter-groups	Crime	<ul style="list-style-type: none"> • EIT Act 2008 <i>jo.</i> 2016 • Criminal Code 1946 • Eliminations of Racial and Ethnic Discrimination Act 2008 • Prevention of Blasphemy Act 1956 • Press Act 1999 • GR ESTI 2019
Contents containing threats of violence or scares aimed toward an individual	Crime	<ul style="list-style-type: none"> • EIT Act 2008 <i>jo.</i> 2016 • Criminal Code 1946 • GR ESTI 2019

Contents containing elements of violence against children	Crime	<ul style="list-style-type: none"> • EIT Act 2008 jo. 2016 • Child Protection Act 2002 jo. 2014 • GR ESTI 2019
Contents containing elements of violations of intellectual property	Crime	<ul style="list-style-type: none"> • Copyrights Act 2014 • EIT Act 2008 jo. 2016 • GR ESTI
Contents containing elements of violations of intellectual property	Crime	<ul style="list-style-type: none"> • Copyrights Act 2014 • EIT Act 2008 jo. 2016 • GR ESTI
Contents containing elements of terrorism and/or radicalism, separatism and/or prohibited dangerous organizations	Crime	<ul style="list-style-type: none"> • Terrorism Act 2003 jo. 2018 • EIT Act 2008 jo. 2016 • GR ESTI 2019
Contents containing elements of violations of trade in goods and services through electronic systems	Crime	<ul style="list-style-type: none"> • EIT Act 2008 jo. 2016 • Trade Act 2014 • GR ESTI 2019
Contents containing elements of violations of information security	Crime	<ul style="list-style-type: none"> • EIT Act 2008 jo. 2016 • State Defence Act 2002 • State Intelligence Act 2011 • Public Information Disclosure Act 2008 • GR ESTI 2019

Contents containing elements of violations of consumer protection	Crime	<ul style="list-style-type: none"> ● EIT Act 2008 jo. 2016 ● Consumer Protection Act 1999 ● GR ESTI 2019
Contents containing elements of violations in the health sector	Crime, Administrative	<ul style="list-style-type: none"> ● Health-related Acts (e.g., Medical Practice Act 2004, Health Act 2009, Hospital Act 2009) ● Narcotics Act 2009 ● EIT Act 2008 jo. 2016 ● GR ESTI 2019
Contents containing elements of violations of supervision of medicine and food	Crime, Administrative	<ul style="list-style-type: none"> ● Health Act 2009 ● Narcotics Act 2009 ● Food Act 2012 ● GR ESTI 2019
Contents that disturb community and public order (e.g., falsified information and/or facts)	Crime	<ul style="list-style-type: none"> ● EIT Act 2008 jo. 2016 ● Criminal Code 1946 ● GR ESTI 2019

Source: compiled by authors, 2021

The table above illustrates that Indonesian laws do not explicitly distinguish what content is 'harmful' and 'illegal'. Consequently, all content listed above could be treated as 'illegal' content, which is substantially a criminal offence. Thus, every person who publishes, disseminates, or distributes content listed above could be criminally punished. It is a logical consequence of the absence of an explicit distinction between illegal and harmful content in Indonesian regulations.

As previously discussed, the Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/66/290) provides guidance to differentiate between illegal and harmful content. Under the Special Rapporteur above, illegal content means content that contains an offense under international law and can be prosecuted criminally, whereas harmful content could be classified as an expression that is not criminally punishable but may justify a restriction or civil suit, and expression that does not give rise to criminal or civil sanctions but still raises concerns in terms of tolerance, civility, and respect for others.

Concern regarding the distinction between illegal and harmful content was also raised in 1996 in the EU through 'Communication on illegal and harmful content on the Internet' (the Communication). This document emphasizes the importance of differentiating online illegal and harmful content. Furthermore, both the Special Rapporteur and the Communication emphasize that different content categories also pose 'radically different issues of principle and call for very different legal and technological responses'.

Hence, the absence of clear distinction between illegal and harmful content in Indonesian regulations potentially poses significant implementation issues in protecting citizens' rights, especially freedom of expression. Moreover, as explained above, there is a higher possibility of criminalizing speech that is not a criminal offense or is not punishable under international standards.

Moreover, the potential legal problems could also be reflected in the various district court decisions presented in each type of content discussed above, such as the District Court of Kebumen in 2018 that broadened the object of defamation to include the honour of a legal entity or state institutions, the District Court of Jayapura in 2020 that broadly interpret the nation of Indonesia as part of 'group' according to Article 28 paragraph (2) of the EIT Act, and the District Court of Kendari that interpret a harsh critique and negative comments toward state institutions as hate

speech under Article 28 paragraph (2) of the EIT Act. These examples could show how provisions concerning harmful content could be interpreted differently. Therefore, at the end, how broad or narrow the classification of harmful content will be will highly depend on how the judiciary interprets the provisions related to harmful content.

Applicable Remedies and Handling Method of Illegal and Harmful Content based on Indonesian Regulations

As aforementioned, Indonesian regulations do not explicitly differentiate between illegal and harmful content, and the methods of handling the various types of content described above are mostly similar. In the EIT Act, all prohibited acts are subject to criminal sanctions. It is the same with the Indonesian Criminal Code, Pornography Act, and Elimination of Racial and Ethnic Discrimination Act. Thus, the publication or dissemination of such content is subject to criminal sanctions and could be criminally prosecuted through criminal court.

However, the laws provide several non-penal court mechanisms and non-court settlement mechanisms for limited purposes: (1) lawsuit for damages caused by non-consent utilization of personal data; (2) lawsuit against electronic system operators for damages caused by them; (3) citizen lawsuit against electronic system operators that caused harm or damage to citizens.⁴⁵ Although several non-penal court mechanisms through civil lawsuits and non-court settlements are provided in the EIT Act, it does not necessarily dismiss the criminal acts committed if the case was already or being prosecuted.

In addition to providing criminal sanctions against offenders, Article 40 paragraph (2a) of the EIT Act also specifies the government's responsibility in preventing prohibited content based on laws and regulations.⁴⁶

⁴⁵See Article 26, paragraph (2), Article 38, and Article 39 of Law No. 11 of 2008 jo. Law No. 19 of 2016 on Electronic Information and Transaction.

⁴⁶See Article 40, paragraph (2a) of Law No. 19 of 2016 on the Amendment of Law No. 11 of 2008 on Electronic Information and Transaction.

Furthermore, according to Article 40 paragraph (2b), the government is authorized to terminate access and/or instruct the electronic system operator to terminate access to content that violates the laws and regulations, which in this case are illegal and harmful content, as listed above. In some instances, the court could order the MOCI to terminate access to such content.

The ESOs (including social media platforms) has an obligation to remove illegal and harmful content based on the request from government institutions. If the social media platforms do not heed this request, according to Article 100 of the GR ESTI, they may be subject to administrative fines, and the government can terminate access to illegal and harmful content unilaterally and even terminate access to the platform. This one-sided mechanism is often problematic as there might be a different interpretation of what content could constitute illegal and harmful between the government and the platforms. Further provisions on the request for termination from the public, other ministries/agencies, police, prosecutor, and the judiciary, as well as the mechanism for termination of access to prohibited contents conducted by the MOCI, and also by electronic service operators are provided in the MOCI Regulation on Electronic Service Operators in Private Sector.

Apart from the aforementioned mechanisms, several Acts also determine specific means based on sectoral regulations. For instance, in the Press Act, the handling mechanism could be done through censorship and ban or restriction. Censorship is a coercive deletion on the part or whole of information materials to be published or broadcast, or warning or notice of intimidation in nature by any party, and/or obligation to report, and acquire permission from the authorized body in conducting journalistic activities.⁴⁸

Whereas, ban or restriction of broadcasting is defined as the discontinuation of publishing and circulation or coercive broadcasting or

⁴⁸See Article 1 number 8 of Law Number 40 of 1999 on Press.

against the law.⁴⁹ The authority of those two remains unclear in the Press Act. However, in practice, the enforcement of harmful content, including the harmful content enacted in the Press Act, is conducted by the MOCI. Furthermore, it is necessary also to refer to other regulations that govern the advertisement placement and other regulations related to the objects/materials/products/services that are being advertised. For instance, the Ministry of Health can ask the MOCI to block cigarette advertising on online media (Ministry of Health of the Republic of Indonesia, 2019). Therefore, it is essential to set a clear and coordinated mechanism in handling press-related content in online media to ensure no functional overlap occurs within the relevant institutions.

Furthermore, in the Public Information Disclosure Act, specific means are provided for violations against content that is held or published by public bodies. Article 19 of the Public Information Disclosure Act states that *'Information and Documentation Management Officers in each Public Body shall carry out the test of consequences as referred to in Article 17 in a meticulous and cautious manner prior to declaring a Public Information as exempted from being accessed by any person.'*

Another Act that specifies a definite mechanism is the Copyrights Act. In processing illegal and harmful content, the Copyrights Act provides means to report copyright infringement in the electronic system to the Minister of Law and Human Rights, and the Minister will verify the report. Furthermore, if there is sufficient evidence of copyright infringement, the Minister will recommend the MOCI block parts or the whole copyright-infringing content in the electronic system or make the services of the electronic system inaccessible.

As explained in the elucidation of Article 56 (1) of the Copyrights Act, the term 'block the content and/or user's access rights can be understood in two ways: first, blocking the content or sites that are providing content services; second, in the form of blocking the access of users to specific sites

⁴⁹See Article 1 number 9 of Law Number 40 of 1999 on Press.

by way of blocking the Internet protocol address or similar. Nevertheless, it is worth noting that there is no clear explanation or measurement on what kind of content will be wholly or partially blocked and what kind of content leads to the inaccessibility of the services of electronic systems. Furthermore, further provisions regarding the electronic system, which is used as a media sharing platform, seem nowhere to be found.

All in all, several key regulations which govern social media platforms in Indonesia may harm online freedom of expression in several ways, such as: forcing the social media platforms to remove 'harmful content' by giving a limited timeframe and imposing sanctions; limiting the methods of content moderation to take down or content removal as the only viable options; lacking the provisions to ensure transparent content moderation practices are conducted both by the government or social media platforms.

The Responsibilities of Social Media Platforms in Regulating Illegal and Harmful Content

Several regulations govern the responsibility of social media platforms—or legally defined as the Electronic System Operator (ESO)—in moderating content in Indonesia. The EIT Act and its implementing regulation—the GR ESTI and the Regulation of Minister of Communication and Informatics No.5 of 2020 (MOCI Regulation 5/20)—lay out the responsibility of platforms in moderating content and the consequences in cases of their noncompliance with the legislation. The EIT Act provides a general overview of how platforms should behave in governing content, and it becomes more detailed—even restrictive in its implementing regulations. However, SAFEnet (2020) reported that, in practice, the EIT Act is usually enacted against individuals (mainly ordinary citizens or part of civil society) by those in the position of power (government officials and businesses).

In GR ESTI, the ESOs or platforms are expected to take down illegal content and content that can ‘disturb community and public order’. In the event of failure, the ESOs are liable for punishment in accordance with the applicable laws. More detailed ESOs’ responsibilities are laid out in the MOCI Regulation 5/20. This regulation contains provisions regarding the time limit given to the ESOs to respond to the government requests for content removals (24 hours after being notified or 4 hours in cases of ‘urgent’ takedown request, without elaborating on which situation is referred to as ‘urgent’). Further, it imposes sanctions on the ESOs who fail to act within the given timeframe. The sanctions range from access blocking to the issuance of fines.

The MOCI Regulation 5/20, a relatively more detailed regulation on content moderation, is problematic for several reasons. For instance, it does not provide any due process on take-down requests, especially those that are made by the government (Article 19, 2021). The short timeframe does not give the ESOs time to assess the content removal request carefully. In the end, it may force the ESOs to comply to avoid administrative punishment—this regulation is punitive in its principle.

The ministerial regulation is presumed to give the government more power to control the information being circulated and to impose censorship on the Internet (Article 19, 2021). It emphasizes take-down or content removal as the only viable option for content moderation. Combined with the short timeframe, the regulation may be used to censor legitimate speech and may threaten democracy in the long run. On a more practical aspect, take-down as the only option of content moderation is not in line with most ESOs’ content moderating mechanisms, especially those used to regulate speech on their platforms.

Although the MOCI Regulation is relatively new, Google reports that Indonesia is in the top 10 of takedown requests and tops the chart in terms of actual content being removed (CNN Indonesia, 2021a). David Graff,

Google's Vice President in Safety and Trust, mentioned that they will try to comply with court orders in respective states. In the case of Indonesia, most of the requests that come from the government are due to national security (CNN Indonesia, 2021a).

In 2019, a MOCI spokesperson would go on to say that Facebook was headstrong when it came to government takedown requests, claiming that they would always have disagreed with the request, citing different interpretations of what constitutes prohibited content (Farras, 2019). In a white paper published by Facebook (elaborated in more detail below), they offered their cooperation while suggesting a more specific government policy and critiquing a takedown-heavy and quantity-driven paradigm in content moderation (Bickert, 2020).

Content takedowns are prevalent insofar as they are requested by the government. Regular users are less heard. In collaborative research by The University of Sydney and The University of Queensland (Sinpeng et al., 2021) (funded by Facebook) they found that individuals are disinclined to report what they deem as violating content, particularly hate speech or other negative content towards minorities and other vulnerable groups, because of their perceived lack of impact on Facebook's moderation practices.

However, while the existing regulations force the platform to comply with takedown requests, they lack the provisions to mandate the transparency of the whole process, both for the government bodies and platforms. Corporate social media platforms often abstract their technological practice and policies; hence, little to no information regarding content moderation reaches the layman. The existing regulations in Indonesia also lack the provisions to mandate platforms or government bodies to disclose the whole process of content moderation practices as per the government's requests. Currently, social media platforms have begun to publish their own transparency report in various forms, but often these reports only focus on the number of cases of 'harmful content' that

were successfully moderated throughout the year instead of how they are moderated. A similar approach is taken by the government—albeit more concerning as they treat deletion of posts and website blocking as milestones. To ensure that content moderation does not compromise freedom of speech, disclosure of numbers is insufficient, and a more detailed and in-depth transparency report is vital.

Sinpeng and others (2021) recommend that Facebook (and other social media companies) be more transparent in their moderation practices (from top to bottom), empower user administrators (not just their own internal or outsourced content moderator), as well as involve third party auditors.

Additionally, although Facebook and other social media companies offer public documents of their moderation policy, they have their own internal document that guides content moderators. It appears that a simplified and user-friendly version of this document is needed to ensure an accountable implementation of content moderation. Moreover, this document should be released in as many regional languages as possible. This can be done by coordinating and working together with local or even organized vulnerable community members. Users should have access to unknown, or hard to find, penalty policies and processes of appeal.

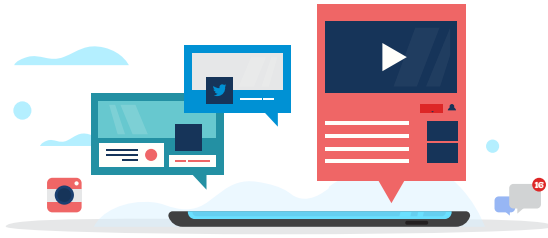




CHAPTER II

Trends





Discussion on the implementation of Indonesian regulation regarding harmful content can be analyzed from various approaches. In this research, two main perspectives are used: first, through the lens of the government and the State as the regulator; and second, through citizens' perspectives as the affected group. The first perspective examines the policies issued by government institutions to respond to content-related issues in online media, while the second perspective highlights various cases that happened in society. Furthermore, issues concerning tech-based and automated content moderation will also be discussed to understand the potential implications of the use of technology in promoting peace and enhancing societies' resilience toward harmful content.

Policies Issued in Responding to Content-Related Problems in Societies

Although many laws and regulations have been enacted, many problems remain as the implementation of the EIT Act and other related regulations still cause multiple interpretations by the police, prosecutor, and judiciary and controversy in the community (See, e.g., Evandio, 2021; Rahmawati, 2021). The government institutions' concern regarding the multi-interpreted implementation of the EIT Act provisions and other related regulations can be seen, for instance, in the latest Joint Decree of MOCI, Chief of National Police, and Attorney General on the Guidelines for the Implementation of Certain Articles in the EIT Act. Therefore, in recent years, state and government institutions have issued many policies related to the implementation of content-related regulations in response to societies' situations. Nevertheless, most of the policies issued are essentially intended as a guide in implementing various regulations explained in the

previous chapter. However, it cannot be neglected that there are certain provisions in these policies that are still prone to create multiple interpretations and harmful to freedom of expression. Elaboration on several relevant policies are provided below.

→ **Circular Letter of The National Police Chief Number SE/6/X/2015 on the Handling of Hate Speech**

The circular letter is aimed to guide the police on hate speech, containing acts to avoid the spread of violence, specifically religious conflicts that happened in 2015. For instance, Tolikara's mosque burning (Jakarta Globe, 2015), and Aceh Singkil's church torching (Dewi Suci Rahayu, 2015) responded to provocative and speculative messages spread online. The circular is essential, especially for those who are part of minority religious groups whose rights have often been violated in the digital space.

The circular defines hate speech by referring to several provisions already existing in the criminal code while broadening the scope of the crime. Hate speech now includes insulting acts (*penghinaan*); defamation (*pencemaran nama baik*); blasphemy (*penistaan*); objectionable acts (*perbuatan tidak menyenangkan*); provocative acts (*memprovokasi*); instigation (*menghasut*), and the dissemination of false news (*penyebaran berita bohong*); and that all of those actions are aimed or affecting to act of discrimination; violence; loss of life; and/or social conflicts.

This circular letter has also differentiated the target of hate speech based on the existing community in such aspects: group of ethnicities (*suku*); religion (*agama*); indigenous religion (*aliran keagamaan*); faith/beliefs (*keyakinan/kepercayaan*); race (*ras*); inter-group relations (*antargolongan*); skin colour (*warna kulit*); ethnicity (*etnis*); gender (*gender*); disability (*kaum difabel (cacat)*); sexual orientation (*orientasi seksual*).

Furthermore, this circular letter also provides a preventive and repressive approach in handling hate speech cases. The preventive approach stresses encouraging police members to understand each form of hate speech, thus enhancing efficient and careful crime prevention. In contrast, the repressive approach focuses on how to enforce the law.

In response to this circular, several activists, media groups, and lawyers have criticized the circular's approach, suggesting that it could lead to human rights violations, namely that of freedom of speech (Mong Palatino, 2015). As mentioned above, this circular letter categorizes insulting acts, defamation, and blasphemy as part of hate speech. However, this could result in wrongful arrest since it does not clearly elaborate the definition and characteristics of those actions. Instead, it refers to provisions in existing regulations, which, as previously mentioned in separate chapters, are broad and can lead to multiple interpretations. The original intention of issuing this circular letter is actually applaudable since the government has started to recognize the existence of hate speech, specifically in the digital space. However, it cannot be neglected that several statements in this circular letter enables subjective decision-making and creates a subjective and legitimate dissent of law enforcement, especially towards the right to freedom of speech and expression.

Telegram Letter of The National Police Chief Number ST/1100/IV/HUK.7.1./2020 4/04/2020 on Violations and Problems That May Occur in the Development of the Situation and Opinions in Cyberspace

The initial background of the issuance of this telegram letter is in response to a large-scale social restriction ('PSBB') during COVID-19 policy implementation (Adi Briantika, 2020). Ever since January 2020, the MOCI has recorded 1.971 hoaxes pertaining to COVID-19 in 5.065 (WM, 2021). This telegram letter is expected to help the government

impose the PSBB policy implementation since misinformation and disinformation related to COVID-19 occur in the digital space. This telegram letter guides the police to conduct a robust cyber patrol to oversee the situations and opinions regarding disinformation about COVID-19, the government's measurement in handling COVID-19, and insults to the President and/or other authorities. Overall, the letter generates the operational rules for police in handling those cases.

This telegram letter orders police investigators to take legal action against any violator regarding the spread of disinformation related to COVID-19. However, several criticisms have been made regarding the issuance of this telegram letter, such as the telegram letter potentially limiting, or even violating, the freedom of expression, opinion, and academic freedom. Lokataru Foundation—the advocacy and movement on human rights and law foundation in Indonesia—mentioned that besides governing the police 'cyber patrol' against COVID-19 'fake news', this telegram letter also regulates the insults against President and/or other Government officials (Lokataru Foundation, 2020). These creates public stigma that rather than to monitor COVID-19 hoax, this telegram letter could be used to silencing and restricting critics to the government (BEM KEMA Universitas Padjadjaran, 2020). The implementation of this telegram is potential to brings about the rise of threats against and arrest of human rights activists for criticising the state's response and policy to COVID-19. Moreover, Amnesty International also proposed to immediately revoke this telegram letter as it is potential to encourage the authorities to abuse their power against the freedom of expression (BEM KEMA Universitas Padjadjaran, 2020).

Moreover, this telegram letter allows police investigators to use the 'COVID-19 emergency' reason to justify their actions in processing the law enforcement since there are no regulations or policies that explicitly and specifically state the measurement and categorization of insult. Therefore, what makes something an insult is also questionable.

As stated above, the response towards the issuance of this telegram letter is that it could lead to violations against the press's rights and public's freedom of speech, especially regarding COVID-19 matters. KontraS, a commission that works to monitor human rights issues, considered the issuance of this telegram letter as a potential threat to the public's right to voice their opinion (especially regarding the COVID-19 issues) and recognized its potential to drive the widespread abuse of power by police officers (Harian Jogja, 2022).

Circular Letter of The National Police Chief Number 2/11/2021 on Ethical Cultural Awareness to Create a Clean, Healthy, and Productive Indonesian Digital Space

This Circular is aimed to elucidate the EIT Act, considering the current regulations did not directly nor specifically explain how to enforce it. In addition, it is also circulated to explain the concern of applying the ITE Law, which relates to the possibility of criminalizing and reporting several parties. Overall, this circular letter can be used to guide the police and prosecutor in handling EIT act violations, including the illegal and harmful content in digital space.

Following the issuance of this circular letter, the Indonesian national police launched a special division, 'virtual police', dedicated to prevention through monitoring, educating, warning, and preventing people from committing a crime in the digital space. The issuance of this circular letter had the same reaction as the previous policy, which recognized that this policy could threaten freedom of expression. As stated by SAFEnet, this might be interpreted as the police's attempt to create a 'digital panopticon'.

The existence of virtual police could potentially lead to technological oppression since it has the power to conduct online censorship, cyber surveillance, and attempts to control an infrastructure (Nugraha and Laila, 2021). According to KontraS, this

circular only governed the establishment of virtual police and the control mechanism, while public information openness regarding the virtual police operation has yet to be regulated. Hence, there is no transparency in the virtual police operation which then made this circular letter open to be interpreted by the authorities. In the first two months of the virtual police operation, there was 329 content that are deemed lawful to EIT Law and 200 of it has passed the verification process (KontraS, 2021). According to KontraS, during the virtual police operation, police took down the content which also contains criticism of the government and the public reprimanded them (KontraS, 2021). Several reports above illustrate that the issuance of this circular letter that becomes the legal basis of the virtual police establishment could also possibly lead to the right to freedom of expression violations (KontraS, 2021).

Telegram Letter of The National Police Chief Number ST/339/II/ RES.1.1.1./2021 22/02/2021 on Guidelines of the Law Management of the Cyber Crime to the Police Investigators

This telegram letter is issued to guide all police investigators in handling cybercrime cases, specifically hate speech. Under these regulations, restorative justice is applicable in criminal proceedings related to the violation against Article 27 (3) of the EIT Act for slander and/or defamation; Article 270 (falsification), 310 (defamation), and 311 (defamation) of the Indonesian Criminal Code. This telegram letter is aimed to help police investigators maximize the law enforcement towards cybercrime perpetrators.

The implementation of this telegram letter can be found in Novel Baswedan's case and Roy Suryo's case. Novel Baswedan was reported by youths, students, and the Community Protection Partnership Student Organization (PPMK) Mitra Kamtibnas over his tweet about the death of Soni Erata alias Ustadz Maaher At-Thuwailibi (VOI, 2021). The Indonesian national police mentioned that in line with what is

stated in this telegram letter, mediation as a form of restorative justice would be used in handling the Novel Baswedan's case (Antara and Kukuh, 2021). Similar to Novel Baswedan's case, Roy Suryo had previously reported Lucky Alamsyah on suspicion of defamation. Lucky Alamsyah was accused of disseminating a hoax in his post, which contained the problem of a traffic accident between him and Roy Suryo (VOI, 2021). This case closed peacefully through mediation.

The issuance of this Telegram Letter is applaudable, since this telegram provides room to prioritize restorative justice while handling several crime in cyber space. This telegram letter encourages the authority to implement restorative justice while handling hate speech case. This also could be used to determine which cybercrime is handled through restorative justice, as it is explicitly stated in the telegram letter which is essential to guarantee legal certainty. As reflected from the aforementioned cases, both prioritized using different approach to handle conflict happened in cyber space—through mediation—that gives affected parties the chance to meet and communicate in order to repair the relationship as well as reducing costly and time-consuming traditional judicial process.

Joint Decree of the Ministry of Communications and Informatics, the Attorney General, and the National Police Chief of the Republic of Indonesia
→ **Number 229 of 2021; Number 154 of 2021; Number KB/2/VI/2021 on the Guidelines for the Implementation of Certain Articles in the EIT Act**

This decree aims to keep a healthy, ethical, and productive Indonesian digital space. This decree will guides the MOCI, the attorney general's office, and the national police in carrying out its duties and authorities to implement the legal enforcement related to violations against the EIT Act. It attempts to unify the interpretation of several provisions related to prohibited content under the EIT Act.

The initial background of this joint decree is the public's concern regarding the rubber articles contained in the EIT Act that are often used for criminalization (Rosana and Eko, 2021). Moreover, several confusing provisions are leading to multiple interpretations by the police, prosecutor, and judiciary. Those original intentions of this joint decree are applaudable. However, this joint decree was criticized because the drafting process was not open and did not involve public participation (Budiarti Utami Putri, 2021). Moreover, the decree is merely an internal guideline for the MOCI, the Attorney General, and the National Police, and does not have a binding power. Koalisi Serius Revisi UU ITE has mentioned that the most crucial thing is rather to amend all of the rubber articles contained in the EIT Act first.

It cannot be neglected that the criminal justice process starts at the police level, which is an entrance to a judicial process. Essentially, all of the aforementioned policies, namely the circular letter, telegram letter, and joint decree, are issued to guide the related institutions, especially the national police, in handling the law enforcement process in harmful content cases, such as hate speech, misinformation, disinformation, and defamation. However, as stated in the beginning of this sub-chapter, there is still room of improvement of several policies, especially to those that often create multiple interpretations between the police, prosecutor, and judiciary.

Trends Through the Lens of Societies

Provisions concerning illegal and harmful content, especially in the EIT Act, have become a matter of public discussion and have been criticized by many civil society organisation ever since the EIT Act was stipulated back in 2008.⁵⁰ The EIT Act has also repeatedly been submitted for constitutional

⁵⁰For instance, Southeast Asia Freedom of Expression Network (SAFE net) has published reports and releases related to EIT Act cases. See reports and release on EIT Act in <https://id.safenet.or.id/?s=UU+ITE>, and list of EIT Act cases in <https://id.safenet.or.id/daftarkasus/>. The list of cases is not updated since December 2020.

review to the Constitutional Court.⁵¹ From all classifications of content provided in the previous chapter, three categories that are most often used as legal grounds for criminal prosecutions are: (1) slander or defamation; (2) hate speech; and (3) content against decency. Moreover, the use of provisions regarding misinformation and disinformation has gradually increased in recent years (See, e.g., Tempo, 2020; Debora, 2020; and SAFEnet, 2021).

Furthermore, throughout 2020, SAFEnet's report on digital rights in Indonesia noted 84 cases of criminalization towards Indonesian internet users, a massive increase compared to the previous year's 24 cases. The EIT Act remains the primary regulation to restrict citizens' speech. The provisions that were used are hate speech (27 cases), defamation (22 cases), and disinformation in electronic transactions (12 cases). Other than the EIT Act, regulations used include the Indonesian Criminal Code, especially Article 14 and Article 15, which regulate misinformation and disinformation, and Arts. 270, 310, and 311, which cover slander and defamation. The discourse remains the same as both regulations, as they are implemented, remain porous to conditions of free speech and expression.

Interestingly, according to reports from Tirto and Detik, the party who used the articles are primarily public officials (See Detik, 2021; Tirto, 2018). Tirto (2018) noted that 35.92% of the people who reported cases on the EIT Act were state officials, including heads of regions, ministerial officers, and police, prosecutor, and the judiciary. In the report, Tirto added that the submission for the EIT Act cases by state or government officials are mainly departed from utterances of expression and criticism of performance or position of officials.

It is confirmed, and, interestingly, increased in 2021, as based on Detik's report, 70% of reporting on the EIT Act to the police from 2017 to 2019 was

⁵¹See e.g., Constitutional Court of the Republic of Indonesia, Constitutional Review Decision, <https://www.mkri.id/index.php?page=web.Putusan&id=1&kat=2&cari=Informasi+dan+Transaksi+Elektronik>

conducted by people with power, including officials, businessmen, and the police themselves. Meanwhile, 29% are carried out by citizens. Moreover, according to SAFENet's report in 2021, out of the 84 subjects reported, 50 are civilians, 15 are activists, four laborers, three private employees, two students, and a journalist. They are victims of allegations of defamation, slander, incitement of intergroup hostility and hate speech, and misinformation/disinformation as they are used to protect and veil government officials and policies.

A coalition of civil society organizations in Indonesia put together a report that urges several reforms to Indonesia's content regulation and content moderation regime (Amnesty International Indonesia et al., 2021). They argue that the current form of content regulation, specifically the EIT Act, and its implementation are frequently on the side of violating freedom of expression, which is far from ensuring peace and national stability. The law is posited to give too much power and discretion to the police without due process and robust accountability measures. Their data showcases those prosecuted with the EIT Act results in a 96.8% conviction rate and 88% incarceration rate; both are deemed to be extremely high.

This coalition that consists of 24 organizations, including Amnesty International Indonesia, Alliance of Independent Journalists, Greenpeace Indonesia, and SAFENet, found that the EIT Act is occasionally used by powerful actors, such as government officials and businessmen, to protect their interests. Notable cases include when Rasio Patra, a researcher, was criminalized using Article 27 paragraph (3) of the EIT Act by Wempy Dyocta Koto, a business motivator, for defamation because of a Facebook post (Widhana, 2017). Wempy is a public figure, frequents social media with business advice, and alleged himself as holding several awards. Sceptical, Rasio researched Wempy's background and found several inconsistencies. For example, Wempy said that he holds Asia's Highest Entrepreneurship Award and Asia's Highest Leadership Award, yet neither award exists. Wempy then clarified that he actually received the Asia Pacific Entrepreneurship Award in 2013 and Asia Corporate Excellence and

Sustainability Award in 2016. Both Ravio and Wempy underwent several threads of correspondence online, with questions and answers flowing between the two of them. But on June 21st, 2017, Ravio was accused of defamation and causing a financial loss for Wempy. Furthermore, Ravio also received outlandish demands such as Rp5 billion in compensation, writing apologies letters to be posted in media outlets, social media, and in video form even though Ravio had never posted a YouTube video.

Moreover, unclear explanations regarding the subject of defamation and the scope of defamatory content also caused polemics. For instance, the Supreme Court in 2010 has held that honour of legal entities, including state institutions, could be an object of defamatory content.⁵² By referring to the 2010 Supreme Court Decision above, in 2018, Syaeful Lillah was held guilty by the District Court of Kebumen of committing defamation against the National Police (Polri) based on Article 27 paragraph (3) of the EIT Act.⁵³ The decision above can illustrate how the EIT Act could be used by governments and state institutions to criminalize citizens.

Another example is when Sadli Saleh, editor in chief of *liputanpersada.com*, wrote an editorial criticizing 'an infelicitous project undertaken by the Central Buton government'. Sadli Saleh was criminalized by Samahuddin, Central Buton's Regent, using Article 28 paragraph (2) that covers incitement of hatred or hostility (Koran Tempo, 2020). Sadli reported that the 5-way junction project ballooned in the budgeting department, from Rp4 billion to Rp6.8 billion, but decreased in lanes from the projected five to four. Furthermore, Sadli also criticised the project's lack of transparency and lacklustre planning, indicated by the ballooning project budget allocation. Irrked by the report, the government official reported the journalist to the authorities for defamation and incitement to hostilities. When in fact, the official could and should have used a right to reply because it concerns a work of journalism (Bernie, 2020). Another example can also be seen in the District Court of Kendari Decision No.

⁵²See Supreme Court Decision No.183 K/Pid/2010 (20 May 2010) 15.

⁵³See District Court of Kebumen Decision No.223/Pid.Sus/2018/PN Kbm (17 December 2018).

426/Pid.Sus/2021/PN Kdi, where the Court held that harsh critique and negative comment toward state institutions could fall under Article 28 paragraph (2) of the EIT Act if that person understands that there will be various responses toward that content, which caused 'social polemic' through social media.⁵⁴ However, there is no clear explanation from the Court regarding what could constitute 'social polemic'.

In 2021, an Indonesian public figure, Jerinx, was also criminalized under Article 28 paragraph (2) of the EIT Act regarding the dissemination of information aimed at causing hatred or hostility to certain individuals and/or community groups based on ethnicity, religion, race, and intergroup.⁵⁵ It started when Jerinx uploaded a post on his Instagram account saying that the Indonesian Doctors Association (IDI) was a 'lackey of the World Health Organization (WHO)'. In its decision, the judge stated that IDI was included in the 'intergroup' group protected by Article 28 paragraph (2) of the EIT Act. However, the judge's decision was deemed inappropriate because the phrase 'intergroup' should not be applied to professional organizations. The Institute for Criminal Justice Reform (ICJR) stated that the equalization of the profession with ethnicity, religion, and race as addressed by Article 28 paragraph (2) of the EIT Act could be dangerous for the democratic climate in Indonesia (ICJR, 2021).

As mentioned before, several courts 'broaden' the interpretation of 'group', to include Presidential and Vice-Presidential Candidate supporters, state institutions (e.g., General Election Commission),⁵⁶ even the nation of Indonesia (*bangsa Indonesia*).⁵⁷ One of the reasons was the constitutional interpretation from Constitutional Court in 2017,⁵⁸ which held that 'intergroup' refers to the sociological reality of the existence of 'other groups' outside ethnicity, religion, and race. Though such flexibility can be

⁵⁴See District Court of Kendari Decision No. 426/Pid.Sus/2021/PN Kdi (16 September 2021) 24.

⁵⁵See District Court of Denpasar Decision No. 72/Pid.Sus/2020/PT.Dps (14 January 2021).

⁵⁶See e.g., District Court of South Jakarta Decision No. 366/Pid.Sus/2019/PN.JKT.SEL (15 August 2019) 84–85.

⁵⁷See District Court of Jayapura Decision No. 16/Pid.Sus/2020/PN Jap (29 April 2020) 38–39.

⁵⁸Constitutional Court Decision No. 76/PUU-XV/2017 concerning Review of Law No. 11 of 2008 on Electronic Information and Transaction as Amended by Law No. 19 of 2016 on the Amendment of Law No. 11 of 2008 on Electronic Information and Transaction (27 March 2018) 66–68.

used to protect groups not explicitly mentioned, such as LGBTQ+ and people with a disability, it is only often used to protect those with sizeable social capital.

Similar to the explanation about the implementation of the provision regarding defamation above, several cases on hate speech above could present an interesting predicament, as those are compounded by an arbitrary notion of a 'group' as regulated in Article 28 paragraph (2) of the EIT Act. With how the terminology could be broadly interpreted, the arbitrary use of the IET Act will restrict citizens' freedom of speech on social media instead of protecting citizens.

Tech-Based and Automated Content Moderation

The relationship between social media users and algorithms is a one-sided affair. On the one hand, platforms and their algorithms know more about the users than they even know of themselves (Reviglio and Agosti, 2019). On the other, veiled by what is called 'software abstraction' where internal details of a software system are intentionally hidden, users hardly know anything about this entity that governs their social media feed (Zulli et al., 2020). This is by no means a limitation of technical ability—not knowing how to code—but also through closed-source code, intellectual property, and copyright restrictions.

Users have little to no idea on how and where their data is being stored, who has access to them, how they are moved across the network, and what conclusions can be drawn from their patterns of interactions. Users are under the watchful eyes of the panopticon. This eye does not only see but also dictates. Karen Yeung (2017) writes on how platforms use big data and algorithms to control not only social media feeds but also the behaviour of users. Yeung (2017, p. 120) posits that platform architecture is often a potent example of a nudge: intentionally designed architecture that alters people's behaviour in predictable ways without actually forbidding other options.

Reviglio and Agosti (2020) see how, when placed in the context of social media algorithms, the specific design choice is one that seeks to persuade—or covertly manipulate—users for the sake of engagement and monetization. They argue that the current state of research into algorithms cannot be done *ex-post* nor *ex-ante* because of the ever-changing technology and the protective nature of platforms. As a suggestion, Reviglio and Agosti propose an audit of platform algorithms to bring more public participation and guarantee democratic oversight. This can come in either the ways of Mastodon and the Fediverse, making the code more open-source or giving access to the public institutions and civil society organizations to help develop, oversee, and develop reports of the algorithm (Zulli et al., 2020).

Similar ideas were brought forth by the Electronic Frontier Foundation’s (2020) policy recommendation for the European Union’s Digital Services Act. In addition to an audit, they suggest that platforms should be more transparent when using algorithms as an automated decision-making process to moderate content. The transparency could be in the form of flagging at which step the algorithm was used, explaining the logic behind the automated process, and explaining how users can contest the decision.

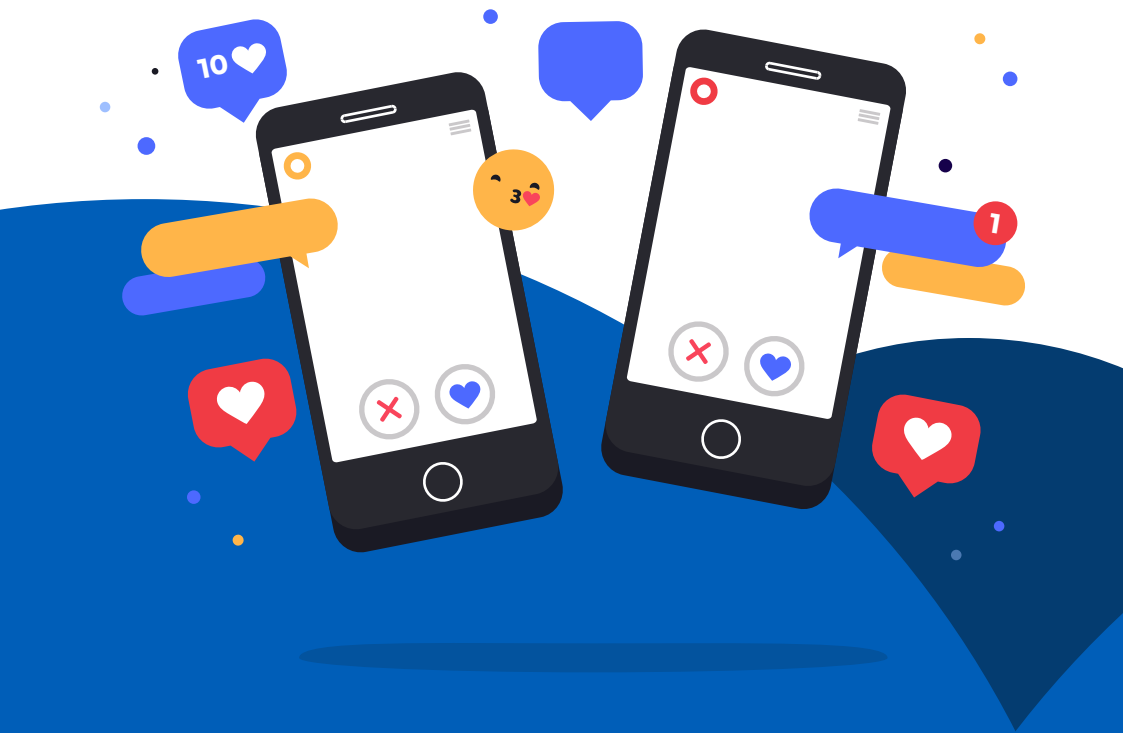
Making algorithms more transparent can also go a long way in changing public perception to demystify the automated processes, mainstreaming the notion that algorithms and AI are not magic and are human-made technology that can also extrapolate, even exacerbate, human problems into the online sphere. Language barriers, resulting in under- or over-flagging of content, is one example (Barrett, 2020; Young, n.d.). This predicament is especially prescient for those living in non-anglophone countries, more so for countries with multiple regional dialects. Of course, there are human moderators to review machine-made decisions, but those machine-made decisions are often the ones that make first contact with the content, and there are a limited number of human content

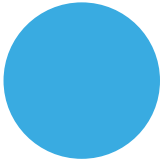
moderators—much less with fair working and living conditions.

Algorithms are also biased, leftover from the biases of their human moderators. Programs are socio-technical products, trained with databases chosen and deemed relevant by the perceptions and knowledge of their creators. Without sensitive considerations for the milieu of social, cultural, and political divergences worldwide, automated content moderation is sure-fire to make ineffective—even harmful—decisions. For example, thousands of videos documenting atrocities during the Syrian Civil War were removed from YouTube, and posts in solidarity with Palestine are flagged as incitement to violence by Facebook (Gorwa et al., 2020; York and Greene, 2021).

Platforms may argue that they have incorporated local knowledge to contextualize their content moderation policy. In non-anglophone countries, the machine is taught to learn from the performance of its local human moderators. However, these human moderators' roles are often limited to implementing the existing guidelines. They are also not necessarily required to have or be equipped with socio-political and cultural awareness on sensitive issues. They also do not necessarily have the medium to give input on the content moderation practices and process (Ahmad, 2018). In short, the participation of these human moderators is not meaningful. Standards made by headquarters that may not be sensitive to the local context will still be enforced regardless. And the AI machine will enforce similar practices, armed with biases from human moderators. This may not be efficient to tackle and combat ongoing misinformation and disinformation in non-English language communities. A report made by Avaaz—an online security group—found that Facebook failed to process 80% of reports on Covid-19 misinformation in Spanish compared to only 27% in English (Valencia, 2021). In Indonesia, there is no available information on how platforms employ these moderators or how many local moderators are employed.

Merylina Lim and Ghadah Alrasheed (2021) suggest the first step in remedying algorithmic bias is to recognize that there is indeed a bias. This might be easy enough to say, but developers often neglect the performativity code and universalize their training. Then, strong commitments must be made, such as diversifying the dataset used for training or even involving developers from otherwise neglected countries. In conjunction with pushing for more transparency and oversight, the biases that algorithms have learned must be unlearned—the same goes for their human counterparts.





CHAPTER III

Concerns



Building on the trends depicted in the previous chapter, this chapter will examine various issues that arise pertaining to harmful and illegal content regulations. This section will then unravel the grey area of numerous matters related to harmful content such as hate speech, misinformation, disinformation, defamation, and the existence of a gap between the platform self-regulatory mechanism with the prevailing regulations and policies enacted by the government. The discussion will also cover the impact of the aforementioned matters in societies, including the impact towards the affected communities that are vulnerable to ill-defined laws.

Regulating Grey Area: Hate Speech, Misinformation, Disinformation, and Defamation.

Most international standards regulating harmful content come in the form of soft law, which is not binding. However, the existing instruments have outlined what the state should do in responding to various matters related to harmful content on the Internet, including hate speech, misinformation, disinformation, and defamation.

International standards regulating hate speech and defamation can be found in the **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on online hate speech (A/74/486)**. It reaffirms that states may restrict freedom of expression in accordance with Article 19(3) of the ICCPR, which requires all regulations to be provided by law and necessary to respect the rights or reputations of others or protect national security, public order, public health, or morals. Furthermore, regulations made by the state must be implemented strictly and supervised transparently. The same instrument also provides an example of a domestic regulation that does not define key terms but imposes significant fines on companies that fail to adhere to its provisions. The regulation is then considered vague, and an unclear definition is deemed to be inconsistent with international human rights law.

In regard to the disinformation and misinformation, the **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on disinformation and freedom of opinion and expression (A/HRC/47/25)** further stipulates that vague laws that confer excessive discretion can lead to arbitrary decision-making and are incompatible with Article 19(3) of the ICCPR. The specific instrument also states that criminalization should be reserved only for the most serious cases.

Reflecting on what has been laid out in the several international standards concerning harmful content, the regulations in Indonesia are far from being in line with those standards. The existing regulations in Indonesia contain several terminologies that have multiple interpretations. This is reflected in several provisions in the EIT Act related to defamation, misinformation, and hate speech. The constitutionality of these provisions has also been tested several times through submission for constitutional review to the Constitutional Court. However, the submissions only result in a bleak outcome. In cases related to defamation, although the MOCI, the Attorney General, and the Chief of the Indonesian National Police have issued guidelines for implementing the EIT Act (Joint Decree), the ambiguity still remains as it did not specify whether an institution could be addressed as an object of defamation. Pertaining to the implementation of the existing regulations and policies, there still occurs criminalization towards civilians in handling defamation, misinformation, and hate speech cases. According to SAFEnet, throughout 2020, the public officials occupy the top positions as reporters of criminalization of the right to freedom of speech in digital space with 47 cases (SAFEnet, 2020). This suggests that the existing regulations and policies are prone to be used by the authority to silence critics and generate unjust persecution.

On the contrary, the existing international standards set out that regulations regarding hate speech, misinformation, disinformation, and defamation shall satisfy the principle of legality that requires the scope, meaning, and effect of the law to be clear, precise, and public. It is further

reaffirmed that vague laws that confer excessive discretion can lead to arbitrary decision-making and are incompatible with Article 19(3) of the ICCPR and therefore, do not meet the international standards that must be fulfilled.

Additionally, the existing regulations also prioritize criminal approach compared to other settlement methods (see Table 3). This is certainly not in line with the international standards that distinguish between content that can be criminalized and content that should fall outside of the scope of the criminal law. Besides, all UN Special Rapporteurs concerning hate speech, misinformation, disinformation, and defamation assert that criminalization should be reserved only for the most serious cases. Therefore, other cases should be addressed by different approaches.

Finally, one can conclude that regulating hate speech, misinformation, disinformation, and defamation in Indonesia is akin to regulating grey areas. Not only because the existing regulations are intertwined and still vague, but the transparency of the implementation of the regulations is also often questioned. In addition, the regulations are not in favour of minorities, and there is a tendency to be used by authorities or people in power.

The Gap Between Platforms' Self-Regulatory Mechanisms and Regulations

It is an established fact that content is governed not solely by states but also by platforms. Self-regulatory mechanisms—or platforms' policies and methods of content moderation—are in place, sometimes in the absence of regulations from domestic governments. Without realizing it, we may have encountered some of these self-regulatory mechanisms. The 'terms and conditions' or the 'terms of services' we accepted when we first made our social media accounts is one self-regulatory tool made by a platform. Another tool, which is the 'community guidelines', is also publicly accessible. The logic for the community guidelines is simple: whenever a

user violates the carefully constructed rules, they may face in-platform consequences, such as account suspension or content removal.

In reality, enforcing a self-regulatory mechanism for content—more familiarly known as content moderation—might pose some challenges for platforms. For one, it is hard for a platform to enforce one applicable standard for content globally (Gillespie, 2018). At times this standard may be too broad for tech-based moderators to understand. For instance, Facebook once removed educational content on breastfeeding for ‘violating’ its community guidelines, misinterpreting the post as pornography (Gillespie, 2018). At other times, this standard—crafted in platforms’ headquarters—may be lost in translation, as many nations’ languages need to be contextually interpreted before they are moderated (Wilson & Land, 2021, p. 1060, p. 1064). However, there are allegations that platforms’ investment in moderating content in a non-English language is severely underfunded. According to a report made by Frances Haugen, a whistle-blower who previously worked at the platform, Facebook spent 87% of its fund to combat misinformation in the English language; meanwhile, only 9% of its users actually speak English (Valencia, 2021). Facebook also admitted that they are used to spread information that fuels the genocide in Myanmar, failing to moderate hate-speech content that ignited the war crime (Stevenson, 2018).

Another significant challenge for platforms in regulating content is accommodating domestic law in their content moderation policy. Platforms and states may have differences in their content moderation approaches. In Indonesia, these differences start from the classification of harmful contents (see Table 2). The differences in the scope of harmful content between platforms and Indonesia’s domestic regulations lead to differences in the handling mechanism. The platform will only remove the content that violates its community guidelines, but they mostly will resort to other, arguably softer, means of moderation when content does not violate the guidelines but may be considered harmful regardless. For instance, platforms may only resort to flagging, labelling, downranking, and

demonetizing harmful content (Audrine & Setiawan, 2021). In contrast, Indonesia treats all harmful content—as classified in various regulations—as illegal. Therefore, content removal is the only method legally recognized to moderate ‘illegal’ content in Indonesia based on the existing regulations.

In a recent white paper published by Facebook (Bickert, 2020), they express their discontent regarding how unclear definitions hinder the moderation process and that publisher liability laws can also have negative consequences such as over- and self-censorship.

**Table 2. Comparison of Content Classification
in Indonesia's Regulation and Platforms' Community Guidelines¹⁹**

Classification of Content in Indonesia's Regulation	Facebook	Instagram	Twitter	YouTube	TIKTOK	WhatsApp	LINE
contents against propriety (including pornography and child pornography)	X	X	X	X	X	X	X
contents of gambling		X			X	X	
contents of slander and/or defamation						X	
contents of extortion and/or threats	X	X	X	X	X	X	
false and misleading information resulting in consumer loss in Electronic Transaction	X			X		X	X
information aimed at inflicting hatred or dissension on individuals and/or certain groups of community-based on ethnic groups, religions, races, and inter-groups	X	X	X	X	X	X	

¹⁹ In this comparison, (x) shows type of contents that are both regulated as ‘illegal’ by the government of Indonesia and classified as ‘harmful’ by the platforms—hence require content moderation mechanism to be enacted against said contents.

Classification of Content in Indonesia's Regulation	Facebook	Instagram	Twitter	YouTube	TIKTOK	WhatsApp	LINE
contain threats of violence or scares aimed toward an individual.	X	X	X	X	X		
Contents containing elements of violence against children	X	X		X	X		X
Contents containing elements of violations of intellectual property	X	X	X	X	X	X	X
Contents containing elements of terrorism and/or radicalism, separatism and/or prohibited dangerous organizations	X	X	X	X	X	X	
Contents containing elements of violations of trade in goods and services through electronic systems	X	X				X	
Contents containing elements of violations of information security	X	X	X	X	X	X	X
Contents containing elements of violations of consumer protection	X					X	
Contents containing elements of violations in the health sector	X	X				X	
Contents containing elements of violations of supervision of medicine and food	X	X				X	
Contents that disturb community and public order (e.g., falsified information and/or facts)							X

Source: compiled by authors, 2021

Platforms are especially careful in regulating speech content, such as in handling misinformation or fake/false news. Twitter, for instance, explicitly states that they are 'serving public conversation' and want to make it easier for users to 'make informed decisions' (Roth & Pickles, 2020), but then decides to label posts that contain misleading information or remove the content based on the severity of the harm the information may cause. Facebook (now Meta) enforces three-part strategies to counter misinformation or false news by demonetizing, downranking, then finally removing the post or the pages that share falsified facts (Meta, 2018). While content removal or account suspension is an option, platforms usually only resort to it after stages of assessment by human platform moderators or AI machines. This assessment process performed by platforms may be improved to be more efficient and precise in handling harmful content. However, this specific due process in regulating content is missing in Indonesia's regulation (Audrine & Setiawan, 2021).

Audrine and Setiawan (2021) posits that the lack of due process leaves platforms with only two extreme choices in moderating content if the government submits a request: to keep or to delete the reported contents. If they fail to comply, platforms are liable for administrative sanctions ranging from fines to access termination. The short timeframe stated in the MOCI Regulation 5/20 and the fear of sanctions may force platforms to remove content without careful assessment and appropriate due process. Additionally, platforms do not have the rights or medium to appeal the decisions made by the government. In many ways, similar to NetzDg, the MOCI Regulation 5/20 shapes Indonesia's content moderation with a punitive approach. This approach may not be the most ideal for content moderation mechanisms that wish to uphold human rights principles, especially freedom of expression. On top of it all, the lack of due process comes without the mechanism that mandates transparency, both to the government and platform. As a result, both the government and platform only release information on the number and types of content they have moderated instead of meaningful transparency reports, which explain how they moderate online content.

The punitive mechanisms and short timeframe to decide and act may incentivize platforms to divert their resources, ceasing proactive moderation, and only wait for reports (Bickert, 2020, pp. 13–14). Moreover, platforms may also prioritize a review of posts reported because they are simply closer to the 24- or 4-hours deadline rather than older (unreported posts), even though that content could also be causing harm. Another metric discussed—one that is seen more favourably in the white paper—is prevalence or the reach and extent of prohibited content, how many are viewing, liking, and sharing it. In this scenario, platforms are incentivized to focus more on one harmful viral content, seen by millions, rather than ten or twenty but are minuscule in terms of engagement. Hence, metrics can be gamified, boosting company resources and attention to those measured and shortcut unmeasured areas. This echoes the concern raised by Alkiviadou (2019) that monitoring cycles correlates with the time-to-action performance of content moderation.

Indeed, Indonesia is currently struggling with issues of negative speech. On the one hand, hate speech and defamation, as it is presently defined, interpreted, and implemented through the Indonesian Criminal Code and the EIT Act produces overcriminalization (Angendari, 2020; Hamid, 2019; Heryanto, 2021; Putri, 2021). On the other hand, those same regulations are unable—or rather the government is unwilling—to protect minority rights, such as religious and gender minorities (George, 2017). However, social media companies, Facebook included, are not exempted from uneven implementation of policy documents (Doctorow, 2021a; Sinpeng & Martin, 2021). In fact, the Electronic Frontier Foundation has an entire project dedicated to documenting these practices (Trendacosta & York, 2019).

Table 3. Comparison between State and Social Media Platforms' Approach in Handling Illegal and Harmful Content in Indonesia

	State	Social Media Platform
Methods	<ul style="list-style-type: none"> • Court: criminal and civil • Non-Court: alternative dispute resolutions, administrative action (including a request for termination of access from the public, the police, prosecutor, judiciary, and other government agencies) 	<ul style="list-style-type: none"> • Content removal • Account suspension • Downranking • Demonetization • Flagging • Label and warning
Subject	<ul style="list-style-type: none"> • Person: court mechanism, a non-court mechanism (especially alternative dispute resolution) • Electronic system operators: non-court mechanism (primarily administrative action), and to some extent could be liable for a criminal offense 	Account/user: restriction (minimum violation) or suspension (multiple violations)
Object	<ul style="list-style-type: none"> • Violations: court mechanism, a non-court mechanism (especially, alternative dispute resolution) • Contents: court mechanism (in case of an order for temporary or permanent injunction), administrative action (takedown request from the government) 	Contents: visibility reduction and demonetization (non-violation but considered harmful); content removal (violation or as per official request)

Source: compiled by authors, 2021.

Impact of Content Regulations Toward Societies

There are a few ways hate speech and misinformation may affect societies. At some point, the two may collide and intertwine with each other. For instance, the dis/misinformation surrounding the 2019 Indonesian election that led to a riot was partly fuelled by Anti-Chinese sentiment (Temby, 2019).

The regulation of speech online affects different communities in different ways, some may be disproportionately more affected than others. In some cases, regulations may be used to those in power to silence the community (as in the cases of politicians using regulations on defamation to prosecute activists and journalists), some communities may not be as protected online due to their identities (the LGBTQ+ community), and to

another community, there is an increased number of cases of hate speech conducted online (the religious minorities.)

→ **Journalists**

Several commentators have warned how the current content regulation regime may have detrimental effects on Indonesian society. The two cases mentioned above accentuate how not only the layman can be affected but also journalists that are supposedly protected by their press privileges. Anton Muhajir (2019), writing for Remotivi, warns how the current laws can create a 'chilling effect' whereby civil society and journalists alike are afraid and censor themselves in fear of being deemed liable for their speech online—which are otherwise valid. Not only in fear of the police and government officials but their fellow citizens. EIT Act has been used widely against journalists that are deemed threatening. For instance, in 2019, the journalist and documentary filmmaker, Dandhy Dwi Laksono, was named as a suspect of hate speech under the EIT Act for his tweet on clashes in Papua (Freiheit, 2021). Muhammad Asrul was sentenced to three months in prison for 'violating' article 45(1) in conjunction with Article 27(3) on hate speech after he wrote articles about an act of corruption done by the son of the Palopo mayor that was published in online media, *beritanews.com* (The Finery Report, 2021).

Furthermore, AJI (Aliansi Jurnalis Independen/Alliance of Independent Journalists) has recorded 15 cases pertaining to journalists' criminalisation for the past three years (CNN Indonesia, 2021b). In 2021 alone, AJI explains, the Indonesian Press Council received 44 reports from the police concerning alleged violations of the EIT Act by journalists (Ningtyas et al., 2021). Even though not all of them went to Court, AJI and the Press Council fears that the trend of using the EIT Act and Criminal Code against journalists is not showing signs of slowing down. In 2015, Indonesia's MOCI released a statement affirming that journalists who produce works of journalism in accordance with the Press Act and journalistic ethics have nothing to worry about (Yura, 2015). MOCI is adamant that the EIT Act have measures that protect journalists. However, on the contrary, numerous

journalists have fallen victim even though they have worked in accordance with the Press Act, journalistic methods, and ethics. This includes Asrul, who was still sentenced to prison even though the judge and the Press Council deemed his work as a credible work of journalism. It has been frequently affirmed that, according to the Press Act, subjects of news articles are entitled to the right to reply, and those news outlets are responsible for taking down or making changes to the article if there are errors in facts. But the rampant use of the EIT Act by government officials and citizens alike has made it the norm.

→ Civil Society Organizations

In addition to journalists, activists often balance their speech, despite otherwise legitimate forms of evaluation and critique. Often helping others who have been wronged by the EIT Act, activists and members of Civil Society Organizations (CSO) themselves must be prudent of government officials who use the EIT Act to stifle criticism and law enforcement who offers their discretion on behalf of the powerful. For example, Haris Azhar, former Coordinator of The Commission for Disappeared and Victims of Violence (KontraS), and Fata Maulidiyanti, current Coordinator of KontraS, were reported by the Indonesian Coordinating Minister of Maritime and Investment Luhut Binsar Pandjaitan for defamation using the EIT Act Article 27 paragraph (3) (Maaruf, 2021). The case stems from a YouTube video on Haris Azhar's channel titled '*ADA LORD LUHUT DIBALIK RELASI EKONOMI-OPS MILITER INTAN JAYA!!JENDERAL BIN JUGA ADA!! NgeHAMtam*' or roughly translated as 'Lord Luhut has economic motive behind the Intan Jaya military operation! The general of Indonesian State Intelligence Agency! NgeHAMtam'. The video discusses a report written by several CSOs that stipulates Pandjaitan's involvement in a mining operation in Intan Jaya, Papua, Indonesia (Tim Advokasi Bersihkan Indonesia, 2021).

Fatia, Haris, and the organizations that compiled the report posit that because of Pandjaitan's position as a public official, a former military officer,

and compounded by the Intan Jaya's heavy military occupation presents a conflict of interest. The Coordinating Minister then issued a subpoena towards Haris and Fatia, constituting what they said as slander and attempting to charge with defamation. However, Nurkholis, Haris's attorney, opined that Pandjaitan's camp has yet to explain which part of the video and report they deem as untrue (Taher, 2021). Furthermore, Asfinawati, Fatia's attorney, stated that Haris and Fatia exercised free expression, had in mind public interests, and targeted Pandjaitan's position as a public official, not as an individual. Hence, Fatia's attorney opines that if the research that Fatia and Haris cited contains erroneous information, the issue should be resolved with a clarification instead of a subpoena. As it stands, the case is developing, and the police are proceeding from a subpoena to a formal investigation (CNN Indonesia, 2022).

→ Citizens

When journalists and activists can be stifled, then the position of the average taxpayer becomes that much more precarious. Take, for instance, WP and AS, two civilians snared by the EIT Act Article 28 paragraph (2) on hate speech (SAFEnet, 2021). WP, a laborer in Riau, was apprehended by the police for allegedly insulting Indonesian President Joko Widodo. He uploaded a meme on Facebook with President Jokowi's face and a quipping remark indicating his discontent with the Indonesian government policies regarding COVID-19. The authorities then charged him with the EIT Act Article 28 paragraph (2) for hate speech and, according to the police, his comments had the potential to incite horizontal conflict (CNN Indonesia, 2020). Meanwhile, AS was also apprehended and charged with the EIT Act Article 27 paragraph (3) for defamation and insulting the Semarang Municipal Government. According to the police, the report against AS came directly from Semarang's Municipal Government, who did not take kindly to AS's Facebook post, which contained aspersions and criticizes their roadblock as COVID-19 containment policy (Rahmayadi, 2020). Both were expressed with the intent to critique and one to humour, and both were said in the context of the government's position in dealing with the

COVID-19 pandemic. In times of crisis, such as the pandemic, both expressions are a catharsis of their discontent with their country's condition and their government's policies. Their expression was far from being unreasonable, much less intended as hate speech.

→ **Marginalized Communities**

Ariel Heryanto (2021), media scholar and Professor of Indonesian Studies at Monash University, Australia, claims that, instead of laws that protect government officials and other powerful groups and individuals from legitimate critique and other forms of speech online, there should be provisions in place to protect the people's dignity and from the officials who may abuse their power and connections. This is compounded by the fact that in the international human rights regime, defamation is no longer stipulated to be contained in criminal law. Instead, it is covered as private law. Heryanto deems incarceration because of critiquing the government—calling it as defamation, inciting hatred, or misinforming the public—as draconian and a colonial legacy that needs to be abolished.

An assemblage of an imperfect EIT Act, unaccountable law enforcement, and officials who may abuse their power are why the Electronic Frontier Foundation and SAFEnet warn against future regulations that hold platforms accountable to the government (Rodriguez, 2021)—calling the MOCI Regulation No. 5 as 'most invasive of human rights'. They suggest reforming the relevant laws before holding platforms accountable to them.

According to research conducted by SAFEnet, various articles were being used to criminalize internet users in 2020 (SAFEnet, 2021). Within 84 cases mentioned above, most cases cite problematic articles of the Law, particularly Article 28(2) on hate speech in 27 cases, Article 27(3) on defamation in 22 cases, and Article 28(1) on consumer loss due to false information in 12 cases. The rise in the amount of criminalization against freedom of expression cannot be separated from the government's handling of the COVID-19 pandemic and the Jobs Creation Omnibus Law

issuance. Cited from SAFEnet, various cases happened to revolve around those two issues.

Regarding the non-court methods, the Ministry of Communications and Informatics, in its 2020 annual report, stated that the total number of access terminations to illegal content on websites stood at 130,254 (Ministry of Communications and Informatics, 2021). Whilst, the total number of access terminations to illegal (and harmful) content on social media platforms reached 183,434. The terminations are implemented using AIS Engine, which is equipped with Artificial Intelligence to search for illegal content in digital space gleaned from the keyword-based search method.

From the total of 130,254 illegal internet content handled and processed, gambling content remained common with 76,216. Approximately 46,172 contents were pornography related, 3,484 contents contained fraud, 2,903 contents were intellectual property rights infringements, 1,366 contents were illegal content reported by sector agencies, 95 contents are violating the information security, ten content contained disinformation, and eight cases contained terrorism, radicalism, sentiment on SARA (tribal, religious, racial and societal group), contents about separatism, contents about dangerous organization, and the rest were not classified. Twitter was found to have the greatest number of illegal contents with 165,698 contents; followed by Facebook, Instagram, and WhatsApp with 5,843 contents; file sharing with 1,272 contents; Google and YouTube with 399 contents; Telegram with 225 contents; and LINE with one content.

In its application, the available content regulations in Indonesia still have many problems, which are quite worrisome. If left unchecked, there will be fears of over-criminalization, especially against civil society.

In the existing regulations, several terminologies have multiple interpretations. This is reflected in several provisions in the EIT Act related to defamation, misinformation on electronic transactions, hate speech, and interception. In fact, the constitutionality of these provisions containing

vague terms has been tested several times through submission for constitutional review to the Constitutional Court.⁶⁰ Of the many requests, only one was granted by the Constitutional Court regarding illegal interception.

In cases related to defamation, the courts in Case No. 223/Pid.Sus/2018/PN Kbm, for instance, refers to the Supreme Court Decision No. 183 K/Pid/2010, which broadens the object of defamation and states that a legal entity could be an object of defamation.⁶¹ Although the MOCI, the Attorney General, and the Chief of the Indonesian National Police issued guidelines for implementing the EIT Act (Joint Decree), it did not specify whether an institution could be addressed as an object of defamation. This ambiguity of meaning then becomes a separate concern in implementing the existing content regulations.

In addition to the existence of multiple interpretations of terminology, the existing regulations also prioritize crime compared to other settlement methods. This is reflected in the mapping of online content classification based on Indonesian regulations in Table 1. Of the various types of content that are regulated, only two types of content allow administrative efforts to be handled. This is certainly not in line with international standards, which distinguish between content that can be criminalized and what cannot.

Apart from problems in content regulation, state and social media platforms often have different opinions regarding illegal and harmful content. This is reflected in the various definitions of illegal and harmful content in states' regulations and social media guidelines. While social media platforms name content identified for flagging or removal as 'harmful content', the state uses the term 'restricted content'. This will lead to different approaches in handling the content in question. Furthermore, there are provisions in the implementing regulations of the EIT Act that are

⁶⁰See Constitutional Court Decision No. 50/PUU-VI/2008 (defamation case), Constitutional Court Decision No. 2/PUU-VII/2009 (defamation case), Constitutional Court Decision No. 52/PUU-XI/2013 (hate speech case), Constitutional Court Decision No. 76/PUU-XV/2017 (hate speech case).

⁶¹See Supreme Court Decision No. 183 K/Pid/2010 (20 May 2010) 15.

not in line with the social media platform guidelines. For example, the existing regulations only adopt a content removal mechanism in dealing with content. Meanwhile, social media platforms already have a variety of content handling mechanisms, such as flagging content that is considered harmful.

Lack of unanimity definitions of illegal and harmful content in Indonesia's legal framework with the social media platform guidelines and the unclarity of each terminology (e.g., morality, public order, etc.) will affect the protection towards freedom of expression. This depicts legal uncertainty, potentially harming marginalized communities such as journalists, activists, etc. As aforementioned, the government and the authorities are subjects that frequently misuse the existing regulations to silence the general public that wants to voice their opinion regarding a particular policy that the government issued. However, the status quo of the stipulated illegal and harmful content provisions has essentially caused governmental repression of criticism and dissent to become 'decentralized' in a way such that there no longer exists national collaboration between the general public and the government, but instead becomes the discretion of rulers in defending their interests (Hamid, 2019).

● **Marginalized Communities: LGBTQ+**

Unfortunately, in terms of protection towards marginalized communities' rights, social media platforms are unsafe and often hostile environments for lesbian, gay, bisexual, transgender, queer, and others (LGBTQ+). The content they post is moderated arbitrarily by the social media platforms that shall follow standards and community guidelines, resulting in many LGBTQ+ accounts, posts, and themed advertisements being taken down through reports and/or ban the procedure. Meanwhile, the homophobic, transphobic, and sexist content often remains untouched (EDRI, 2019). LGBTQ+ community for sure has faced certain hardships due to those double standards.

Despite that, Indonesia has not explicitly criminalised LGBTQ+ for expressing identities and sexual practices in social media under its national regulations. Indeed, there are certain limitations, such as what is being regulated in Article 4 (1) and its elucidation of the Pornography Act, that prohibit the creation, dissemination, or broadcasting of pornography containing sexual intercourse, including lesbian sex and male homosexual sex. However, the Indonesian constitution has stipulated the right of all Indonesians, presumably also including those that are part of the LGBTQ+ community—even though it has not been explicitly mentioned—to be free from discriminatory treatment.

On the contrary, there are various cases pertaining to discrimination and the limitation of the right to freedom of expression and opinions. For instance, to commemorate pride month, the Indonesian Journalists Union for Diversity has created a digital space for community solidarity through a live webinar in June 2020 (New Mandala, 2020). Regrettably, the hopes to promulgate important information regarding LGBTQ+ rights and protection were crushed when YouTube, in response to viewer complaints concerning the sensitivity of the content, removed the webinar mid-broadcast. Furthermore, gay activist Hartoyo overcame virtual curbs in his movement. When Jakarta entered a lockdown in April, the offline exhibition as a part of financing aid programs through the in-house SriKendes boutique, which sells original, traditional motif clothing, has switched to a virtual exhibition. Hartoyo began tinkering with Facebook, under the account name of 'Har Toyo' and eventually settled on a new fundraising model by auctioning pre-loved clothes, handbags, and shoes on Facebook Live. However, the users began reporting his account, leading to two Facebook suspensions. The reason why Facebook Indonesia has suspended his account was because of the usage of the word *bencong*—an Indonesian slur for 'queer'. That word was classified by Facebook algorithms as hate speech (Norman Harsono, 2020). Moreover, the Ministry of Informatics and Communications even restricts LGBTQ+ content by taking down

the websites (See Ministry of Communication and Information Technology, 2018a; Ministry of Communication and Information Technology, 2016; Ministry of Communication and Information Technology, 2018b).

Sinpeng et al. (2021) observed that there was an ominous presence of hate speech in the comment section of three LGBTQ+ communities' pages on Facebook. The negative comments on each Facebook page represent at least one-third of all public comments. There has been an increased number of hateful comments and hate speech ever since the then Minister of Technology, Research, and Higher Education, suggested that LGBTQ+ students should be banned from campus to protect the morals and norms in Indonesia. Through the interview with the admins of respective LGBTQ+ community pages, the hateful comments and hate speech are not only common on Facebook but also other platforms, such as Twitter, Youtube, and Instagram. Indeed, in some cases the administrators themselves have been attacked and harassed through the private messaging channel (Sinpeng et al., 2021).

The acceptance and tolerance of the LGBTQ+ community in Indonesia cannot be separated from socio-cultural attitudes and religious values. Acceptance means that they can participate in all family and social life without reservations, whereas tolerance is usually expressed grudgingly or out of necessity (UNDP & USAID, 2014). Religion has become an essential part of constructing social and legal norms in Indonesia. Generally speaking, all of the religions that are accepted in Indonesia are against the LGBTQ+ practice (Andina, 2016). In line with that, the majority of Indonesia's socio-cultural values and norms have also opposed the existence of LGBTQ+; however, there are several changes in terms of rigidity, and the culture is becoming more open towards diversity.

● **Marginalized communities: Religious Minorities**

In addition to the LGBTQ+ community, one of the marginalized groups that have become the target of hate speech in Indonesia is the Ahmadiyya community—an Islamic sect frequently thought to tarnish Islam, Indonesia’s majority religion (CSIS Indonesia, 2021; Solikhin, 2016). Based on data from the Center for Strategic and International Studies (CSIS) Indonesia, the number of acts of hate speech against the Ahmadiyya group as a religious minority group on social media, namely Twitter, from 1 January 2019 to 31 July 2021 reached an alarming number. In 2008, Sobri Lubis, one of the leaders of the Islamic Defenders Front (FPI), appeared in a video on YouTube, lecturing while shouting, 'Kill! Kill! Kill! Kill the Ahmadis!' (George, 2017). Alisa Wahid, General Coordinator of the Jaringan Gusdurian, said that the provocation of hate speech on social media was then often used to commit acts of violence in the real world (Apriyani, 2017).

In 2016, a riot in Tanjung Balai, North Sumatra, was incited by a chain of messages through word of mouth and social media (The Jakarta Post, 2016). The message was spun, distorted from a conversation mentioning how a mosque’s speaker increased in volume through the years to allegations of a non-Muslim intent to prohibit azan (call to prayer) (Pusat Studi Agama dan Demokrasi Paramadina, 2018). A mob of agitated Muslims would go on to ransack one Buddhist temple and three pagodas, destroying prayer equipment, tables, chairs, cars, motorbikes, statues, and plundered the houses of worship in the process.

Similar events happened in Singkil, Aceh and Tolikara, Papua. Inter-religious conflict, which resulted in a church being burned in Singkil, was attributed to hate speech being spread on social media (Lubabah, 2015). An Indonesian atheist was arrested in charge with blasphemy for posting on Facebook regarding his belief and disbelief

in his previous religion, Islam (Cochrane, 2015). Anti-Shia content is also prevalent in Indonesian social media (Azali, 2017). Many more episodes can be found regarding the intersections of religion and the power imbalance produced by a majority-minority divide (us and them mentality).

Islamophobia—hate speech and fear of Islam—in social media has been rising in Indonesia. Influenced by an external factor, orientalist visions of Islam from the west, and three internal factors (Kastolani, 2021): first, a reaction against intolerant sermons from fundamentalist religious figures; second, a form of freedom of expression; and third, an accumulation of Indonesian internet citizens' identity politics and religious polarization.

Panjaitan (2016) suggests that a permissive government and a low literacy rate of Indonesian internet users intertwined to produce an unsafe, unpeaceful environment for religious minorities on social media. Although Indonesia already possesses regulations against hate speech, both online and offline, it remains relatively high. Panjaitan offers two possible causes. First, it is because the public is not aware of these regulations. Second, it is compounded by the lack of precedent of government enforcement of these laws, particularly to protect religious minorities.

However, Panjaitan (2015) also found that banning hate speech as problematic. Historically, the Indonesian government—even after the transition from authoritarianism to democracy—has used laws to silence and criminalize those who would criticize the government. For Panjaitan, the solution should be thought of as more in the societal realm, which is increasing digital literacy such as filtering before sharing (*saring sebelum sharing*).



CHAPTER IV

Conclusions and Recommendations



Conclusions

As the number of social media users is increasing from year to year, efforts to regulate illegal and harmful content on the Internet are crucial. This is also because of the phenomenon that the rising circulation of content is not accompanied by high digital literacy score. Given the complexity and the ramifications of the issue, this research shows that efforts to regulate illegal and harmful content are not the sole responsibility of the government but also the private sector, such as social media platforms. Based on the analysis of the existing national legal framework, social media platforms' self-regulation, and its implementation, this research highlights three important notes regarding illegal and harmful content regulations and its implementation in Indonesia as elaborated below.

→ Content Classification and Definition

The findings, as can be seen in Table 1, echo that Indonesian regulations and policies did not differentiate between 'harmful' and 'illegal' content. Almost all harmful content is classified as illegal and therefore as criminal acts. Thus, the publication or dissemination of such content is subject to criminal sanctions. Therefore, it can be concluded that content regulation in Indonesia is not yet in accordance with international standards which have tried to distinguish the 'harmful' and 'illegal' content.

In the international realm, even binding legal instrument that distinguishes between illegal and harmful content are also rarely found. Mostly, the distinction is contained in a 'soft' form that is not legally binding. However, the instrument can still be used as a reference. It is strengthened by the importance of differentiating 'harmful' and 'illegal' content in regulating content on social media as there may be differences in handling and duty of care for illegal and harmful content in practice between states and social media platforms.

Therefore, there is an urgency to adopt international standards on categorizing illegal and harmful content. The standards are then expected to be a benchmark for states to be able to handle illegal and harmful content in a more targeted manner.

Additionally, there is also a jarring mismatch between the government and the platform's classification of harmful content. This mismatch may result in differences in handling content online, thus reducing the efficiency of platforms' content moderation. Platforms are also the subject of domestic content regulation, where they may face the dilemma of compromising their self-regulatory mechanisms or being punished for compliance failure.

→ Handling Mechanism

From Table 2, it can be seen that each stakeholder has a different approach in dealing with illegal and harmful content. Through its regulations and policies, the state has two methods in dealing with illegal and harmful content in Indonesia, namely court and non-court. Court methods include handling cases through criminal and civil courts. In contrast, non-court methods include alternative dispute resolutions and administrative actions such as termination of content and platforms, administrative fines, etc.

In Indonesia, the platform's handling mechanism may significantly differ from the State's. The only methods of moderation legally recognized in Indonesia are content removal and access blocking. While platforms have several methods to use, namely: downranking, demonetizing, flagging, labelling, and warning, removing content, and suspending accounts. Platforms are also especially careful with speech content. They may refuse to remove the content of false news but may resort to reducing its visibility. With MOCI Regulation 5/20, platforms may be liable for punishment if they refuse to remove content, resulting in a punitive approach to Indonesia's content regulation. This may further threaten the rights of citizens as it may result in censorship, self-censorship, and other acts that may violate fundamental rights.

→ Remaining Problems

Although state regulations and policies—as well as social media platforms guidelines—are in place and have provided various mechanisms in dealing with illegal and harmful content, problems to effectively regulate illegal and harmful content on the Internet remain.

1

First, the state's regulations and policies—especially relating to hate speech, misinformation, disinformation, and defamation—are still not in line with existing international standards. While clarity became the primary requirement in making regulations according to international standards, there are normative problems such as vague terms that can be broadly interpreted in existing state's regulations and policies and no explicit distinction between illegal and harmful content as previously mentioned.

2

Second, there are implementation problems related to subjective and different interpretations by the authorities, including government officers, police, prosecutor, the judiciary, and even the public. This problem is reflected by the substance discrepancy in many national policies issued and considerations of district court decisions as explained above. As a result, the existing issues also have detrimental effects on marginalized communities, such as the LGBTQ+ community and religious minorities.

3

Third, the problem lies in the relationship between the state and social media platforms in handling illegal and harmful content. In some cases, there is still disagreement between the state and social media platforms in handling the content. This is reflected in the different definitions of illegal and harmful content in the state regulations and social media guidelines.

Fourth, social media platforms need to make improvements in

4

their self-regulatory mechanisms. Platforms need to invest more in non-English language content moderation and make the content moderation process more transparent. While platforms have stated their commitment to improving content moderation practices, data shows that this initiative may still be lacking, especially to combat misinformation online. Local content moderators still do not have meaningful participation in moderating local content, and the existing transparency report still does not possess information on the moderating process.

5

Fifth, there are still no provisions that mandate meaningful transparency in Indonesia's content moderation. While both the government and platforms release information on the types and numbers of content they have moderated, they do not provide detailed information on how they do it, including how they decide that content is harmful. This condition, combined with the lack of due process, of the content moderation practices in Indonesia may harm the freedom of expression online.

Nevertheless, with the limited timeframe and resources, the research was carried out mostly by analyzing secondary data. Therefore, many things concerning the implementation of content regulation can still be developed and explored empirically, for instance, regarding the impact on marginalized communities. Furthermore, this research will be the baseline for Phase II activities of the SM4P project by setting up a national multi-stakeholder coalition in Indonesia, developing an online monitoring mechanism for potentially harmful content, and organising consultations with the conflict-affected communities to understand the impact of harmful content on conflict dynamic and increasing act of intolerance against religious, gender, and sexual minorities in Indonesia. This research is also expected to encourage further discussion and research on the impact of content regulation for society and marginalized community groups in Indonesia. Finally, through various findings and recommendations

explained in this research, as well as activities during this project that bring together multiple stakeholders, the overall SM4P project is expected to gather different thoughts and concerns from various stakeholders to stimulate better changes in regulating content and handling harmful content for the Indonesian government and social media platforms.

Recommendations

The results of the legal mapping and analysis echo that there is still a lot of room to be improved to increase the effectiveness of the regulation of illegal and harmful online content in Indonesia. As aforementioned, it is worth noting that the step towards reform is not the sole responsibility of the state. Thus, commitment from the private sector, such as social media platforms, is also critically needed.

Recommendation 1:



Revising the EIT Act and its implementing regulations

The EIT Act serves as the primary legal basis regulating cyberspace in Indonesia. Therefore, the EIT Act also has implementing regulations, such as GR ESTI and MOCI Regulation 5/20, which have also been described in the legal mapping. However, in its implementation, several areas need to be improved, both in the EIT Act and its implementing regulations. These include:

Reformulating the regulation of content classification

This research recommends that the state reformulate content types in the EIT Act and its implementing regulations. Courts often use a conventional approach, which can lead to over-criminalization. Reformulation can be carried out by adopting international standards. Presently, several international legal instruments have provided a basis for distinguishing between illegal and harmful content—or at least mapping content that can be criminalized and not. For example, the Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and

expression (A/66/290) distinguishes three types of content, namely: (a) expression that constitutes an offense under international law and can be prosecuted criminally; (b) expression that is not criminally punishable but may justify a restriction and a civil suit; and (c) expression that does not give rise to criminal or civil sanctions, but still raises concerns in terms of tolerance, civility, and respect for others.

● **Redefining terms related to illegal and harmful content**

This research recommends the State redefine terms related to illegal and harmful content. In some cases, judges, police, and prosecutor use terms that have a broad meaning in handling cases deemed illegal and harmful. For example, cases of mis/disinformation are handled using articles related to defamation. In the online context, there are new terms that have their own definitions and cannot necessarily be interpreted with existing terms, such as misinformation and disinformation.

● **Reforming the content handling mechanism**

This research recommends the State reform the current content handling mechanism in the EIT Act and its implementing regulations. The update includes several things. First, the regulation shall accommodate a variety of content handling mechanisms. Currently, the existing regulations only adopt a content removal mechanism. Meanwhile, social media platforms already have a variety of content handling mechanisms, such as flagging content that is considered harmful. *Second*, the regulation shall add time for social media platforms to handle the flagged content. The current regulations have very rigid terms of time, even though social media platforms need to check content before taking it down. *Third*, the regulation shall reduce sanctions relief for social media platforms. The high number of sanctions in the regulations now burdens platforms and makes them create mechanisms just to circumvent the sanctions. *Fourth*, the regulation shall add an appeal mechanism for social media platforms.

Fifth, reformulating the sanction for harmful content-related cases. As mentioned above, most of the sanctions toward harmful content-related cases are criminal sanctions. However, legal measures through criminal sanction toward harmful content should be the last resort. Other comprehensive approaches through education and technological means could be employed.

Recommendation 2:

→ Harmonizing the laws and regulations related to illegal and harmful content

This research recommends the state harmonize laws and regulations related to illegal and harmful content. The harmonization is projected to reduce the possibility of different interpretations and overlaps between regulations. This recommendation is not only carried out on the regulations governing the online realm but also on other regulations that may intersect with the issue of illegal and harmful content.

Recommendation 3:

→ Equalizing perceptions of the meaning of the provisions of actions prohibited in the EIT Act and other related Acts

This research recommends the State provide common perceptions of the meaning of the provisions of actions prohibited in the EIT Act and other related Acts. In some cases, an act can be interpreted differently by the police, prosecutor, and judiciary. Therefore, there must be a unified perception of the provisions of prohibited acts. This can be achieved by making a joint decree or other legal instruments. Furthermore, in hate speech, mis/disinformation, and defamation cases, the police, prosecutor, and judiciary could also adopt the six-part threshold test based on the Rabat Plan of Action in determining whether the content is harmful, namely: (a) context, (b) speaker, (c) intent, (d) content and form, (e) extent of the speech act, and (f) likelihood. The threshold test could be incorporated into regulations, such as the EIT Act.

Recommendation 4:

- Enhancing cooperation between the State and social media platforms in handling illegal and harmful content

This research recommends that the state and social media platforms could enhance their cooperation, as, in handling harmful content, there should be a shared responsibility between state and social media platforms. In this case, it must be ensured that the state and social media platforms are moving in the same direction in dealing with illegal and harmful online content, especially regarding hate speech, mis/disinformation, and defamation content that has increased in recent years. Therefore, various discussions and multi-stakeholder meetings need to be encouraged.

Recommendation 5:

- Increasing transparency in moderating content

This research recommends that both the state and social media platforms ensure meaningful transparency on implementing their content regulation. Meaningful transparency means that they do not stop creating an output-based report (e.g., numbers of content being moderated) but also inform the citizens over the whole process of content moderation.



Bibliography

- Ahmad, S. (2018). 'It's Just the Job': Investigating the Influence of Culture in India's Commercial Content Moderation Industry.
- Alkiviadou, N. (2019). Hate speech on social media networks: Towards a regulatory framework? *Information & Communications Technology Law*, 28(1), 19–35. <https://doi.org/10.1080/13600834.2018.1494417>.
- Amnesty International Indonesia, et al. (2021), Kertas Kebijakan: Catatan dan Desakan Masyarakat Sipil atas Revisi UU ITE.
- Andina. (2016). Faktor Psikososial dalam Interaksi Masyarakat dengan Gerakan LGBT di Indonesia, *Aspirasi: Jurnal Masalah-masalah Sosial* 7(2), 173–185. <http://jurnal.dpr.go.id/index.php/aspirasi/article/view/1288/709>.
- Angendari, D. A. D. (2020, November 30). Definisi 'ujaran kebencian' di Indonesia terlalu luas, gampang dimanfaatkan. *The Conversation*. <http://theconversation.com/definisi-ujaran-kebencian-di-indonesia-terlalu-luas-gampang-dimanfaatkan-150743>.
- Anindyajati, T. (2021). Limitation of the Right to Freedom of Speech on The Indonesian Constitutional Court Consideration. *Indonesian Law Journal*, 14(1), 19–36. <https://doi.org/10.33331/ilj.v14i1.45>.
- Antara and Kukuh S. Wibowo (2021). Terapkan Surat Edaran Kapolri Soal UU ITE, Kasus Novel Baswedan akan Dimediasi. <https://nasional.tempo.co/read/1435856/terapkan-surat-edaran-kapolri-soal-uu-ite-kasus-novel-baswedan-akan-dimediasi> (accessed 2 February 2022).
- Apriyani, R. (2017). Gusdurian: Provokasi Ujaran Kebencian di Medsos Dorong Kekerasan di Dunia Nyata. *KBR*. https://kbr.id/nasional/02-2017/gusdurian__provokasi_ujaran_kebencian_di_medsos_dorong_kekerasan_di_dunia_nyata/88912.html.
- Article 19. (2004). Briefing Note on International and Comparative Defamation Standards.
- Article 19. (2017). Germany: The Act to Improve Enforcement of the Law in Social Networks.
- Article 19. (2021). Malaysia: Emergency (Essential Powers) (No. 2) Ordinance 2021 (Fake News Ordinance).
- Article 19. (2021a). Indonesia: Regulation of the Minister of Communication and Informatics Number 5 of 2020 on Private Electronic System Operators (Ministerial Regulation 5)

- Aschoff, N. (2020, May 29). Social Media Companies Can't Be Trusted to Protect Our Democracy. *Jacobin*. <https://jacobinmag.com/2020/05/donald-trump-twitter-social-media-companies>.
- ASEAN Declaration to Prevent and Combat Cybercrime 2017.
- ASEAN Framework and Joint Declaration to Minimize the Harmful Effects of Fake News 2018.
- Association for Progressive Communication (APC). (2018). Content Regulation in the Digital Age. Submission to the United Nations Special Rapporteur on the Right to Freedom of Opinion and Expression.
- Audrin, P & Setiawan, I. (2021). Impact of Indonesia's Content Moderation Regulation on Freedom of Expression. *CIPS Policy Paper (38)*. <https://repository.cips-indonesia.org/media/347642-impact-of-indonesias-content-moderation-88ec21cc.pdf>.
- Azali, K. (2017). Fake News and Increased Persecution in Indonesia. *ISEAS Perspective*, 61, 10.
- Banchik, A.V. (2021). Disappearing Acts: Content Moderation and Emergent Practices to Preserve at-risk Human Rights Related Content. *New Media & Society* 23(6), 1527–1544.
- Banko, M., MacKeen, B., and Ray, L. A Unified Typology of Harmful Content, Proceedings of the Fourth Workshop on Online Abuse and Harms, 20 November 2020, <https://aclanthology.org/2020.alw-1.16.pdf>
- Barrett, P. (2020). Who Moderates the Social Media Giants? A Call to End Outsourcing. NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/tech-content-moderation-june-2020?_ga=2.264803983.1095278896.1633969975-1758350172.1632663756.
- Benedek, W., & Kettemann, M. C. (2013). Freedom of Expression and the Internet. Council of Europe. https://book.coe.int/fr/attachment?id_attachment=709
- BEM KEMA Universitas Padjadjaran (2020). Kajian Overcriminalization Telegram. <https://kema.unpad.ac.id/wp-content/uploads/kajian-overcriminalization-telegram.pdf>
- Bernie, M. (2020, February 3). Cara Murahan Sikapi Berita: 2 Jurnalis Sulawesi Dijerat Pasal Karet. *tirto.id*. <https://tirto.id/cara-murahan-sikapi-berita-2-jurnalis-sulawesi-dijerat-pasal-karet-ewsR>
- Bickert, M. (2020). Charting a Way Forward on Online Content Regulation. Facebook. https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.
- Branden, B., Davidse, S., and Smit, E. (2021). In between illegal and harmful: A look at the Community Guidelines and Terms of Use of online platforms in the light of

- the DSA proposal and the fundamental right to freedom of expression. DSA Observatory. <https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/>.
- Briantika, Adi (2020). Lima Telegram Kapolri jadi Pedoman Tangani Pelanggaran PSBB, <https://tirto.id/lima-telegram-kapolri-jadi-pedoman-tangani-pelanggaran-psbb-eLk4> (accessed 2 February 2022).
- Budapest Convention on Cybercrime 2001.
- Burns, H. (2020). Online Harms: Freedom of Expression Remains Under Threat. Open Rights Group. <https://www.openrightsgroup.org/blog/online-harms-freedom-of-expression-remains-under-threat/>.
- CNN Indonesia. (2020, April 8). Diduga Hina Jokowi soal Corona, Buruh di Kepri Ditangkap. <https://www.cnnindonesia.com/nasional/20200408192303-12-491818/diduga-hina-jokowi-soal-corona-buruh-di-kepri-ditangkap>.
- CNN Indonesia. (2022, January 6). Kasus Luhut Vs Haris Azhar Naik Penyidikan di Polda Metro Jaya. CNN Indonesia. <https://www.cnnindonesia.com/nasional/20220106160339-12-743459/kasus-luhut-vs-haris-azhar-naik-penyidikan-di-polda-metro-jaya>.
- CNN Indonesia. (2021a). Google: Indonesia Jadi Negara Terbanyak Minta Hapus Konten. *teknologi*. Retrieved February 3, 2022, from <https://www.cnnindonesia.com/teknologi/20211025150358-185-712049/google-indonesia-jadi-negara-terbanyak-minta-hapus-konten>.
- CNN Indonesia. (2021b). AJI: 3 Tahun Terakhir Ada 15 Jurnalis Media yang Dijerat UU ITE. <https://www.cnnindonesia.com/nasional/20211201151322-12-728525/aji-3-tahun-terakhir-ada-15-jurnalis-media-yang-dijerat-uu-ite> (accessed 7 February 2022).
- Cochrane, J. (2014, June 3). Embrace of Atheism Put an Indonesian in Prison. *The New York Times*. https://www.nytimes.com/2014/05/04/world/asia/indonesian-who-embraced-atheism-landed-in-prison.html?_r=0.
- Commission of the European Communities. (1996). Illegal and harmful content on the Internet – Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of the Regions. Brussels, 16 October 1996.
- Constitutional Court of the Republic of Indonesia. (2010). Naskah Komprehensif

- Perubahan UUD 1945 Buku VIII. Secretariat General and Registrar of the Constitutional Court.
- Constitutional Court Decision No. 50/PUU-VI/2008 on the Review of Law No. 11 of 2008 on Electronic Information and Transaction against the 1945 Constitution of the Republic of Indonesia.
- Constitutional Court Decision No. 52/PUU-XI/2013 Concerning Review of Law No. 11 of 2008 on Electronic Information and Transaction against the 1945 Constitution of the Republic of Indonesia.
- Constitutional Court Decision No. 76/PUU-XV/2017 concerning Review of Law No. 11 of 2008 on Electronic Information and Transaction as Amended by Law No. 19 of 2016 on the Amendment of Law No. 11 of 2008 on Electronic Information and Transaction.
- Council of the European Union. (2014). EU Human Rights Guidelines on Freedom of Expression Online and Offline.
- CSIS Indonesia. (2021). CSIS National Hate Speech Dashboard. CSIS Indonesia. <https://dashboard.csis.or.id/hatespeech/>
- De Gregorio, G. (2020). Democratizing Online Content Moderation: A Constitutional Framework. *Computer Law & Security Review* 36, 105374.
- De Streef, A. (2020). Online Platforms' Moderation of Illegal Content Online. Study requested by the European Parliament's Committee on Internal Market and Consumer Protection.
- Debora, Y. Daftar Pasal UU ITE yang Sering Menjerat Netizen di Medsos, <https://tirto.id/daftar-pasal-uu-ite-yang-sering-menjerat-netizen-di-medsos-gbdg>.
- Detik (2021) Para Penunggang UU ITE, <https://news.detik.com/x/detail/investigasi/20210301/Para-Penunggang-UU-ITE/> (accessed 16 October 2021).
- Dimas Jarot Bayu. (2020). Indonesia Negara yang Paling Banyak Blokir Iklan Daring. <https://databoks.katadata.co.id/datapublish/2021/03/03/indonesia-negara-yang-paling-banyak-blokir-iklan-daring> (accessed 30 November 2021).
- District Court of Bale Bandung Decision No. 84/Pid.Sus/2021/PN.Blb.
- District Court of Bandung Decision No. 471/Pid.Sus/2020/PN.Bdg.
- District Court of Bantul Decision No. 125/Pid.Sus/2018/PN.Btl.
- District Court of Denpasar Decision No. 72/Pid.Sus/2020/PT.Dps.
- District Court of East Jakarta Decision No. 532/Pid.Sus/2020/PN.Jkt.Tim.

- District Court of Jantho Decision No.76/Pid.Sus/2021/PN Jth.
- District Court of Jayapura Decision No.16/Pid.Sus/2020/PN Jap.
- District Court of Kebumen Decision No.223/Pid.Sus/2018/PN Kbm.
- District Court of Kendari Decision No.426/Pid.Sus/2021/PN Kdi.
- District Court of Merauke Decision No.132/PID.B/2010/PN.MRK.
- District Court of North Jakarta Decision No.1537/Pid.B/2016/PN JKT.UTR.
- District Court of South Jakarta Decision No.203/Pid.Sus/2019/PN.Jkt.Sel.
- District Court of South Jakarta Decision No.366/Pid.Sus/2019/PN.JKT.SEL.
- Doctorow, C. (2021a, July 16). Right or Left, You Should Be Worried About Big Tech Censorship. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2021/07/right-or-left-you-should-be-worried-about-big-tech-censorship>.
- Doctorow, C. (2021b, August 3). With Great Power Comes Great Responsibility: Platforms Want To Be Utilities, Self-Govern Like Empires. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2021/08/utilities-governed-empires>,
- EDRI. (2019). The Digital Rights of LGBTQ+ People: When Technology Reinforces Societal Oppressions. European Digital Rights. <https://edri.org/our-work/the-digital-rights-lgbtq-technology-reinforces-societal-oppressions/> (accessed 10 December 2021).
- European Parliament. (2020). Report on Digital Services Act and fundamental rights issues.
- European Union Communication on Illegal and Harmful Content on the Internet 1996.
- Evandio, A. CIPS: Revisi UU ITE Diperlukan, Masih Banyak Pasal Multitafsir. Bisnis. <https://kabar24.bisnis.com/read/20210908/15/1439555/cips-revisi-uu-ite-diperlukan-masih-banyak-pasal-multitafsir> (accessed 2 December 2021).
- Evans, D. G. (2009). Human Rights and State Fragility: Conceptual Foundations and Strategic Directions for State-Building. *Journal of Human Rights Practice*, 1(2), 181–207.
- Farras, B. (2019). Astaga! Facebook Paling 'Bandel' Tak Ikuti Arahan Pemerintah. CNBC Indonesia. Retrieved February 3, 2022, from <https://www.cnbcindonesia.com/tech/20190514094646-37-72267/astaga-facebook-paling-bandel-tak-ikuti-arahan-pemerintah>.
- Flew, T., Suzor, N., & Martin, F. (2019). Internet Regulation as Media Policy: Rethinking the Question of Digital Communication Platform Governance. *Journal of Digital Media Policy*, 10(1), 33–50.

- Freiheit. (2021). Press Freedom: Indonesia Government Fears on Critics and Murals. Friedrich Naumann Foundation.org. <https://www.freiheit.org/indonesia/press-freedom-indonesia-government-fears-critics-and-murals>(accessed 7 February 2022).
- George, C. (2017). Indonesia: Demokrasi yang Diuji di Tengah Intoleransi. Pelintiran Kebencian: Rekayasa Ketersinggungan Agama dan Ancamannya bagi Demokrasi (pp. 145–178). PUSAD Paramadina and IIS UGM. <https://www.paramadina-pusad.or.id/buku/pelintiran-kebencian-rekayasa-ketersinggungan-agama-dan-ancamannya-bagi-demokrasi/>.
- Gorwa, R. (2019). The Governance Triangle: Conceptualising the Informal Regulation of Online Content. *Internet Policy Review* 8(2), 1–22.
- Gorwa, R., Binns, R., & Katzenbach, K. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7 (1).
- Government Regulation No. 71 of 2019 on Electronic System and Transaction Implementation.
- Gelber, K. (2019). Differentiating Hate Speech: A Systemic Discrimination Approach. *Critical Review of International Social and Political Philosophy*, 24(4). Retrieved from: <https://doi.org/10.1080/13698230.2019.1576006>.
- Hamid, U. (2019, November 20). UU ITE dan merosotnya kebebasan berekspresi individu di Indonesia. *The Conversation*. <http://theconversation.com/uu-ite-dan-merosotnya-kebebasan-berekspresi-individu-di-indonesia-126043>.
- Hamid, U. (2019, November 25). Indonesia's Information Law has Threatened Free Speech for More than a Decade. This Must Stop. <https://theconversation.com/indonesias-information-law-has-threatened-free-speech-for-more-than-a-decade-this-must-stop-127446> (accessed 3 December 2021).
- Harian Jogja (2022). KontraS Beberkan Fakta Demokrasi Memburuk Begini Cara Pengusaha Membungkam Masyarakat Sipil, <https://news.harianjogja.com/read/2022/01/06/500/1092729/kontras-beberkan-fakta-demokrasi-memburuk-begini-cara-pengusaha-membungkam-masyarakat-sipil>(accessed 2 February 2022).
- Harsono, N. (2020). In Conservative Indonesia, This Gay Activist Braves Social Curbs to Help the Marginalized. *The Jakarta Post*. <https://www.thejakartapost.com/news/2020/08/25/in-conservative->

- indonesia-this-gay-activist-braves-social-curbs-to-help-the-marginalized.html (accessed 10 December 2021).
- Hootsuite. (2021). Digital 2021, Global Overview Report. https://hootsuite.widen.net/s/zcdrtxwczn/digital2021_globalreport_en
- Heryanto, A. (2021, June 26). Gila Hormat. Kompas.id. <https://www.kompas.id/baca/opini/2021/06/26/gila-hormat-2/>.
- Indonesian Criminal Code.
- Indonesian Institute of the Independent Judiciary. (2018). Penafsiran terhadap Pasal 156A KUHP tentang Penodaan Agama (Analisis Hukum dan Hak Asasi Manusia). Policy Brief.
- Institute for Criminal Justice Reform (ICJR). (2021, January 19). Putusan Banding Jerinx: Hakim Gagal Koreksi Pertimbangan yang dapat Berujung pada Malapetaka di Indonesia. <https://icjr.or.id/putusan-banding-jerinx-hakim-gagal-koreksi-pertimbangan-yang-dapat-berujung-pada-malapetaka-di-indonesia/> (accessed 6 February 2022).
- International Covenant on Civil and Political Rights (ICCPR) 1976.
- International Covenant on Economic, Social and Cultural Rights (ICESCR) 1976.
- International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) 1969.
- Jakarta Globe (2015, July 22) Police to Name Suspects in Tolikara Arson Case: Chief. <https://jakartaglobe.id/news/police-name-suspects-tolikara-arson-case-chief/> (accessed 2 February 2022)
- Joint Decree of Minister of Communications and Informatics, the Attorney General, and the Chief of the Indonesian National Police No. 229 of 2021, No. 154 of 2021, No. KB/2/VI/2021.
- Kastolani. (2020). Understanding the delivery of Islamophobic hate speech via social media in Indonesia. *Indonesian Journal of Islam and Muslim Societies*, 10(2), 247–270. <https://doi.org/10.18326/ijims.v10i2.247-270>
- Katadata. (2020). Puncak Penyebaran Hoaks Terjadi Menjelang Pilpres 2019: Jumlah Temuan Hoaks (2019). *Katadata.com*. <https://databoks.katadata.co.id/datapublish/2020/01/07/jelang-pemilu-hoaks-makin-berseliweran> (accessed 3 December 2021).
- Keen, C., Kramer, R., & France, A. (2020). The Pornographic State: The Changing Nature of State Regulation in Addressing Illegal and Harmful Content Online. *Media Culture and Society* 42(7–8), 1175–1192.
- Khan, I. (2021). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on disinformation and freedom of expression (A/HRC/47/25).

- Koalisi Tolak Kriminalisasi dan Rekayasa Kasus. (2020, April 23). [PERNYATAAN BERSAMA KOALISI TOLAK KRIMINALISASI DAN REKAYASA KASUS] Segera Lepaskan Rasio Patra, Hentikan Kriminalisasi, Ungkap Pelaku Peretasan! Institute for Criminal Justice Reform. <https://icjr.or.id/pernyataan-bersama-koalisi-tolak-kriminalisasi-dan-rekayasa-kasus-segera-lepaskan-rasio-patra-hentikan-kriminalisasi-ungkap-pelaku-peretasan/>.
- KontraS (2021) Pemutakhiran Data Virtual Police, <https://kontras.org/2021/04/22/pemutakhiran-data-virtual-police/> (accessed 2 February 2022).
- Koran Tempo. (2020, February 10). Stop Kriminalisasi Sadli. Tempo. <https://kolom.tempo.co/read/1305497/stop-kriminalisasi-sadli>
- La Rue, F. (2011). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/66/290).
- Law No. 8 of 1999 on Consumer Protection.
- Law No. 40 of 1999 on Press.
- Law No. 11 of 2008 on Electronic Information and Transaction.
- Law No. 19 of 2016 on the Amendment of Law No. 11 of 2008 on Electronic Information and Transaction.
- Law No. 28 of 2014 on Copyrights.
- Lokataru Foundation (2020). Shrinking Civic Space Amidst the COVID-19. <https://www.icnl.org/wp-content/uploads/Lokataru-Foundation-SCS-and-COVID19-Pandemic-Indonesia.pdf>
- Lubabah, R. (2015, November 2). Edaran hate speech disebabkan kasus Tolikara dan Aceh Singkil. Merdeka.Com. <https://www.merdeka.com/peristiwa/edaran-hate-speech-disebabkan-kasus-tolikara-dan-aceh-singkil.html>
- Lundahl, O. (2020). Algorithmic meta-capital: Bourdieusian analysis of social power through algorithms in media consumption. *Information, Communication & Society*, 1–16. <https://doi.org/10.1080/1369118X.2020.1864006>
- Madiega, T. (2020, May). Reform of the EU liability regime for online intermediaries. In-depth Analysis of European Parliamentary Research Service.
- Marcetic, B. (2021, October 9). Facebook Harms Its Users Because That's Where Its Profits Are. <https://jacobinmag.com/2021/10/facebook-whistleblower-haugen-profits-addiction-public-utility-regulation-censorship-moderation-zuckerberg>.

- Meta. (2018, May 23rd). Hard Questions: What's Facebook's Strategy for Stopping False News? <https://about.fb.com/news/2018/05/hard-questions-false-news/>.
- Ministry of Communication and Informatics. (2016). Kemenkominfo: Stiker LGBT LINE Tidak Berlaku di Indonesia. [Kominfo.go.id. https://kominfo.go.id/content/detail/6730/kemenkominfo-stiker-lgbt-line-tidak-berlaku-di-indonesia/0/sorotan_media](https://kominfo.go.id/content/detail/6730/kemenkominfo-stiker-lgbt-line-tidak-berlaku-di-indonesia/0/sorotan_media) (accessed 10 December 2021).
- Ministry of Communication and Informatics. (2017). Berantas Iklan Kesehatan Hoaks, Kemenkes Gandeng 7 Lembaga. [Kominfo.go.id. https://kominfo.go.id/index.php/content/detail/12121/berantas-iklan-kesehatan-hoaks-kemenkes-gandeng-7-lembaga/0/sorotan_media](https://kominfo.go.id/index.php/content/detail/12121/berantas-iklan-kesehatan-hoaks-kemenkes-gandeng-7-lembaga/0/sorotan_media) (accessed 30 November 2021).
- Ministry of Communication and Informatics. (2018a). Ini Dia Upaya Gigih Menkominfo Babat Habis LGBT. [Kominfo.go.id. https://www.kominfo.go.id/content/detail/12448/ini-dia-upaya-gigih-menkominfo-babat-habis-lgbt/0/sorotan_media](https://www.kominfo.go.id/content/detail/12448/ini-dia-upaya-gigih-menkominfo-babat-habis-lgbt/0/sorotan_media) (accessed 10 December 2021).
- Ministry of Communication and Informatics. (2018b). Penanganan Konten yang Melanggar Nilai dan Norma Sosial Budaya. [Kominfo.go.id. https://kominfo.go.id/index.php/content/detail/12403/siaran-pers-no08hmkominfo012018-tentang-penanganan-konten-yang-melanggar-nilai-dan-norma-sosial-budaya/0/siaran_pers](https://kominfo.go.id/index.php/content/detail/12403/siaran-pers-no08hmkominfo012018-tentang-penanganan-konten-yang-melanggar-nilai-dan-norma-sosial-budaya/0/siaran_pers) (accessed 10 December 2021).
- Ministry of Communication and Informatics (2019) Siaran Pers No. 67/HM/KOMINFO/03/2019. [Kominfo.go.id. https://www.kominfo.go.id/content/detail/17465/siaran-pers-no-67hmkominfo032019-tentang-kominfo-pantau-dan-siap-blokir-iklan-kampanye-di-platform-digital-selama-masa-tenang/0/siaran_pers](https://www.kominfo.go.id/content/detail/17465/siaran-pers-no-67hmkominfo032019-tentang-kominfo-pantau-dan-siap-blokir-iklan-kampanye-di-platform-digital-selama-masa-tenang/0/siaran_pers) (accessed 30 November 2021).
- Ministry of Communication and Informatics. (2021). Annual Report Ministry of Communications and Information [Kominfo.go.id. https://www.kominfo.go.id/content/detail/36485/annual-report-ministry-of-communication-and-information-technology/0/laporan_tahunan](https://www.kominfo.go.id/content/detail/36485/annual-report-ministry-of-communication-and-information-technology/0/laporan_tahunan) (accessed 3 December 2021).
- Ministry of Communication and Informatics. (2021). Siaran Pers No. 143/HM/KOMINFO/04/2021.

- https://kominfo.go.id/content/detail/34136/siaran-pers-no-143hmkominfo042021-tentang-sejak-2018-kominfo-tangani-3640-ujaran-kebencian-berbasis-sara-di-ruang-digital/0/siaran_pers (accessed 3 December 2021).
- Ministry of Health of the Republic of Indonesia (2019). Kemenkes Meminta Kemkominfo Blokir Iklan Rokok di Internet. P2PTM.Kemkes.go.id. <http://p2ptm.kemkes.go.id/kegiatan-p2ptm/pusat-/kemenkes-meminta-kemkominfo-blokir-iklan-rokok-di-internet> (accessed 16 October 2021).
- Morozov, E. (2019, February 4). Capitalism's New Clothes. *The Baffler*. <https://thebaffler.com/latest/capitalisms-new-clothes-morozov>.
- Muhajir, A. (2019). Kenapa Publik Dirugikan Kalau Jurnalis Dijerat UU ITE? Retrieved December 1, 2021, from <https://www.remotivi.or.id/amanat/516/kenapa-publik-dirugikan-kalau-jurnalis-dijerat-uu-ite>.
- Newell, K. (2020). LGBTQ+ Community Leaders in Indonesia: Overcoming Pandemic Hardship. *New Mandala*. <https://www.newmandala.org/lgbtq-community-leaders-in-indonesia-overcoming-pandemic-hardship/> (accessed 10 December 2021).
- Nickel, J. W. (1993). How Human Rights Generate Duties to Protect and Provide. *Human Rights Quarterly*, 15(1).
- Ningtyas, I., Musdalifah, Faisal, E., Maryadi, O., & Afrida, N. (2021). Catatan Akhir Tahun 2021: Kekerasan, Kriminalisasi & Dampak UU Cipta Kerja (Masih Bayangi Jurnalis Indonesia). *Aliansi Jurnalis Independen (AJI)*. https://aji.or.id/upload/article_doc/Catahu_AJI_2021.pdf.
- Nugraha, Ricky Mohammad and Laila Afifa (2021). State Uses Virtual Police for Mass Surveillance, SAFENet Says. <https://en.tempco.co/read/1439061/state-uses-virtual-police-for-mass-surveillance-safenet-says> (accessed 2 February 2022).
- Oates, S. (2020). The easy weaponization of social media: Why profit has trumped security for U.S. companies. *Digital War*, 1(1–3), 117–122. <https://doi.org/10.1057/s42984-020-00012-z>.
- Palatino, M. (2015, November 29). Will Indonesia's Police Circular on Hate Speech Suppress Freedom of Expression? <https://advox.globalvoices.org/2015/11/29/will-indonesias-police-circular-on-hate-speech-suppress-freedom-of-expression/#>
- Panjaitan, R. P. (2016, August 18). Mob violence shows Indonesia must act against

- Regional Regulation of Special Region of Yogyakarta No. 2 of 2017 on Peace, Public Order and Community Protection.
- Regional Regulation of Buton Regency No. 2 of 2020 on the Implementation of Peace, Peace, Public Order and Community Protection.
- Reviglio, U., & Agosti, C. (2020). Thinking Outside the Black-Box: The Case for 'Algorithmic Sovereignty' in Social Media. *Social Media + Society*, 6(2), 205630512091561. <https://doi.org/10.1177/2056305120915613>.
- Rodriguez, K. (2021, February 16). Indonesia's Proposed Online Intermediary Regulation May be the Most Repressive Yet. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2021/02/indonesias-proposed-online-intermediary-regulation-may-be-most-repressive-yet>.
- Rosana, Fransisca Christy and Eko Ari Wibowo (2021, June 25). 5 Fakta Seputar Terbitnya SKB Pedoman UU ITE, <https://nasional.tempo.co/read/1476302/5-fakta-seputar-terbitnya-skb-pedoman-uu-ite/full&view=ok> (accessed 2 February 2022).
- Roth, Y & Pickles, N. (2020). Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information
- SAFEnet (2020, April 6). Pejabat Publik Paling Banyak Laporkan Kasus Pidana UU ITE. <https://databoks.katadata.co.id/datapublish/2021/04/06/safenet-pejabat-publik-paling-banyak-laporkan-kasus-pidana-uu-ite> (accessed 2 February 2022).
- SAFEnet (2021). Laporan Situasi Hak-hak Digital Indonesia 2020, Represi Digital di Tengah Pandemi. <https://id.safenet.or.id/2021/04/laporan-situasi-hak-hak-digital-indonesia-tahun-2020-represi-digital-di-tengah-pandemi/>
- Sander, B. (2021). Democratic Disruption in the Age of Social Media: Between Marketized and Structural Conceptions of Human Rights Law. *European Journal of International Law*, 32(1), 159–194.
- Setianti, L. and Djafar, W. (2017). Tata Kelola Konten Internet di Indonesia: Kebijakan, Praktik, dan Permasalahannya. Policy Brief.
- Setiawan, I. (2021). Who is Responsible for User-Generated Content on Digital Platforms in Indonesia? Policy Brief(8).
- Sinpeng, A., & Martin, F. R. (2021, July 5). Facebook's failure to pay attention to non-English languages is allowing hate speech to flourish. The Conversation. <http://theconversation.com/facebooks-failure-to-pay-attention-to-non-english-languages-is-allowing-hate-speech-to-flourish-163723>
- Sinpeng, A., Martin, F. R., Gelber, K., & Shields, K. (2021). Facebook: Regulating

- Hate Speech in the Asia Pacific. The University of Sydney and The University of Queensland.
https://r2pasiapacific.org/files/7099/2021_Facebook_hate_speech_Asia_report.pdf
- Solikhin, A. (2016). Islam, Negara, dan Perlindungan Hak-Hak Islam Minoritas. *Journal of Governance*, 1(1).
- Stevenson, A. (2018, November 6th). Facebook Admits It Was Used to Incite Violence in Myanmar. *New York Times*.
<https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html>
- Supreme Court Decision No.183 K/Pid/2010.
- Susanto, E. (2013). Undang-Undang Keterbukaan Informasi Publik dan Penyelenggaraan Pemerintahan. *Komunikator*, 5(1), 54–58.
- Taher, A. P. (2021, September 22). Haris Azhar & Fatia Dipolisikan Luhut, Kuasa Hukum: Salah Alamat. *Tirto.id*. <https://tirto.id/haris-azhar-fatia-dipolisikan-luhut-kuasa-hukum-salah-alamat-gjK7>.
- Temby, Q. (2019). Disinformation, Violence, and Anti-Chinese Sentiment in Indonesia's 2019 Election. ISEAS Yusof of Ishak Institute Perspective Report, (129). Retrieved from:
https://www.iseas.edu.sg/images/pdf/ISEAS_Perspective_2019_67.pdf.
- Tempo. ITE Law's Malleable Terms – The Tale of Suppressive Twins,
<https://interaktif.tempo.co/proyek/pasal-karet-uu-ite-sejoli-pembungkam-kritik/index.php>.
- The Finery Report. (2021). Journalist from Palopo Sentenced to Prison after Investigation Corruption Case.
<https://www.thefineryreport.com/news/2021/11/26/journalist-from-palopo-sentenced-to-prison-after-investigating-corruption-case>
 (accessed 7 February 2022).
- Tim Advokasi Bersihkan Indonesia. (2021, September 23). Haris Azhar dan Fatia Maulidiyanti Dilaporkan Luhut Binsar Panjaitan, Ancaman Serius Terhadap Demokrasi dan Kerja-Kerja Pembela Hak Asasi Manusia. *KontraS*.
<https://kontras.org/2021/09/23/haris-azhar-dan-fatia-maulidiyanti-dilaporkan-luhut-binsar-panjaitan-ancaman-serius-terhadap-demokrasi-dan-kerja-kerja-pembela-hak-asasi-manusia/>
- Tirto. (2018). 'Jerat UU ITE Banyak Dipakai oleh Pejabat Negara',
<https://tirto.id/jerat-uu-ite-banyak-dipakai-oleh-pejabat-negara-c7sk>
 (accessed 30 November 2021).

- The 1945 Constitution of the Republic of Indonesia.
- The Jakarta Post. (2016, July 31). False news spread on social media incited riot in N. Sumatra: Police chief. The Jakarta Post. <https://www.thejakartapost.com/news/2016/07/31/false-news-spread-on-social-media-incited-riot-in-n-sumatra-police-chief.html>.
- Trendacosta, K., & York, J. (2019, May 20). Tossed Out: Highlighting the Effects of Content Rules Online. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2019/05/tossed-out-highlighting-effects-content-rules-online>.
- United Nations Educational, Scientific and Cultural Organization. (2018). Module 2: Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. Available at: https://en.unesco.org/sites/default/files/f._ifnd_handbook_module_2.pdf.
- United Nations Development Programme & USAID. (2014). Being LGBT In Asia: Indonesia Country Report. <https://www.usaid.gov/documents/2496/being-lgbt-asia-indonesia-country-report-bahasa-language> (accessed 10 December 2021).
- United Nations High Commissioner for Human Rights. (2013). Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred.
- United Nations Strategy and Plan of Action on Hate Speech 2019.
- Universal Declaration of Human Rights 1948.
- Valencia, S. (2021, October 28th). Misinformation online is bad in English. But it's far worse in Spanish. <https://www.washingtonpost.com/outlook/2021/10/28/misinformation-spanish-facebook-social-media/>
- Vogelezang, F. (2020, December 2). Illegal vs Harmful Online Content. <https://www.internetjustsociety.org/illegal-vs-harmful-online-content>
- VOI (2021, February 15). After Being Polished, Novel Baswedan was Reported to the KPK. <https://voi.id/en/news/33130/after-being-polished-novel-baswedan-was-reported-to-the-kpk> (accessed 2 February 2022)
- VOI (2021, September 20). Roy Suryo Reveals Mazdjo Pray's Identity, Convicted of Guilty at the Tangerang District Court 2018 Ago. <https://voi.id/en/news/86732/roy-suryo-reveals-mazdjo-prays-identity-was-convicted-of-guilty-at-the-tangerang-district-court-2018-ago>

- (accessed 2 February 2022).
- Widhana, D. (2017, October 13). Bagaimana Ravio Dilaporkan Wempy lewat UU ITE. *tirto.id*. <https://tirto.id/bagaimana-ravio-dilaporkan-wempy-lewat-uu-ite-cyi8>
- Wilson, R.A. & Land, M.K. (2021). Hate Speech on Social Media: Content Moderation in Context. *Connecticut Law Review*, 52(3)
- WM, (2021). Kemkominfo: 1.971 Hoax COVID-19 Sejak Januari 2020, <https://www.beritasatu.com/nasional/849683/kemkominfo-1971-hoax-covid19-sejak-januari-2020> (accessed 2 February 2022).
- World Economic Forum. (2021). *Advancing a Digital Safety: A framework to Align Global Action*.
- Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, 20(1), 118–136. <https://doi.org/10.1080/1369118X.2016.1186713>
- York, J., & Greene, D. (2021, May 25). Amid Systemic Censorship of Palestinian Voices, Facebook Owes Users Transparency. *Electronic Frontier Foundation*. <https://www.eff.org/deeplinks/2021/05/amid-systemic-censorship-palestinian-voices-facebook-owes-users-transparency>.
- Young, H. (n.d.). The digital language divide. *The Guardian*. Retrieved December 3, 2021, from <http://labs.theguardian.com/digital-language-divide/>.
- Zulli, D., Liu, M., & Gehl, R. (2020). Rethinking the 'social' in 'social media': Insights into topology, abstraction, and scale on the Mastodon social network. *New Media & Society*, 22(7), 1188–1205. <https://doi.org/10.1177/1461444820912533>.





UNIVERSITAS
GADJAH MADA



CENTER FOR DIGITAL SOCIETY



Funded by the
European Union



Center for Digital Society


Faculty of Social and Political Sciences
Universitas Gadjah Mada
Room BC 201-202, BC Building 2nd Floor,
Jalan Socio Yustisia 1
Bulaksumur, Yogyakarta, 55281, Indonesia

Phone : (0274) 563362, Ext. 116
Email : cfds.fisipol@ugm.ac.id
Website : cfds.fisipol.ugm.ac.id

 facebook.com/cfdsugm

 Center for Digital Society (CfDS)

 [cfds_ugm](https://www.instagram.com/cfds_ugm)

 [@cfds_ugm](https://wa.me/cfds_ugm)

 [@cfds_ugm](https://twitter.com/cfds_ugm)

 CfDS UGM