ISEAS YUSOF ISHAK INSTITUTE

ISSUE: 2024 No. 51

ISSN 2335-6677

PERSPECTIVE

RESEARCHERS AT ISEAS – YUSOF ISHAK INSTITUTE ANALYSE CURRENT EVENTS

Singapore | 11 July 2024

Holding Social Media Companies Accountable for Enabling Hate and Disinformation

*Nuurrianti Jalli**

People are seen looking at mobile phones ahead of Malaysia's 15th general election in Wilayah Persekutuan, Malaysia, which was held on 13 November 2022. Election campaigns are changing to attract the attention of individuals through digital propaganda. (Photo by Syaiful Redzuan/ANADOLU AGENCY/Anadolu Agency via AFP).

**Nuurrianti Jalli is Assistant Professor of Strategic Communications at Oklahoma State University and a Visiting Research Fellow at ISEAS – Yusof Ishak Institute, Singapore.*

EXECUTIVE SUMMARY

- Social media platforms, while connecting billions and amplifying marginalised voices, have become tools for spreading hate, disinformation, and extremist ideologies due to these business models prioritising engagement and ad revenue.

- Engagement-driven algorithms incentivise the spread of harmful content, since inflammatory and divisive posts often garner the most attention, creating a cycle that prioritises profits over societal well-being.

- Social media companies have often been seen as hesitant to enforce their policies against misleading political ads due to the substantial revenue these ads generate. The challenge is further compounded by the high cost of effective content moderation in non-English languages, which creates additional barriers to maintaining platform integrity.

- Voluntary self-regulation by social media companies has been inadequate. Governments and international organisations need to step in to enforce meaningful standards for content moderation. Potential approaches include substantial fines for repeated failures, mandatory investment in content moderation, regular third-party audits, and re-examination of legal frameworks to hold companies accountable for algorithmic amplification of harmful content.

- The power of social media companies, if unchecked, poses a danger to democratic institutions. The failure to moderate online content can fuel real-world violence, deepen societal divisions, and erode public trust in democracy. Coordinated regional and global efforts are crucial to ensure consistent and effective standards for social media governance.

INTRODUCTION

The rise of social media has transformed the way we communicate, share information, and engage with the world around us. Platforms like TikTok, Facebook, and X have connected billions of people across the globe, enabling unprecedented levels of interaction and exchange. These platforms have given a voice to the voiceless, empowered activists and marginalised communities, and facilitated the spread of knowledge at an unparalleled scale.

However, as these platforms have grown in size and influence, they have also become powerful tools for spreading hate, disinformation, and extremist ideologies. The business models of these companies, which prioritise user engagement and advertising revenue above all else, have created a toxic online environment that is having devastating real-world consequences. From fuelling ethnic violence and political polarisation to undermining public trust in democratic institutions, the unchecked power of social media giants poses an existential threat to our societies.

PLATFORMS APPROVING PROBLEMATIC CONTENTS

The real-world impact of social media companies' failure to effectively moderate their platforms can be seen in recent examples from India and Malaysia. In India, an investigation by India Civil Watch International (ICWI)[1] found that Meta had approved a series of political ads containing blatant anti-Muslim hate speech, conspiracy theories targeting opposition leaders, and calls for violence during the recent election on its platform, Facebook. These AI-manipulated ads featured slogans like "let's burn this vermin" and false claims that an opposition leader wanted to "erase Hindus from India." Despite Meta's public commitment to crack down on hate speech and disinformation, these ads were allowed to run and reach millions of users. This scandal is not an isolated incident for Meta. In 2022, a similar issue arose in Norway, where political ads containing far-right content and misinformation[2] were discovered on the platform. More recently, a report by The Bureau of Investigative Journalism[3] revealed that for several months in 2024, more than 8,000 ads featuring AI-manipulated videos and false information about politicians had circulated on Facebook. These incidents highlight a persistent problem with Meta's content moderation practices, which have consistently failed to prevent the spread of hateful and misleading content during sensitive political events.

In Southeast Asia, a similar lack of urgency in moderating problematic content has been observed. During the 2022 Malaysian election, TikTok, a platform owned by ByteDance, became a hotbed for inflammatory content promoting an ultra-Malay nationalist agenda. Posts and videos calling for a repeat of the tragic "May 13" racial riots of 1969,[4] which claimed hundreds of lives, gained traction on the platform. These provocative messages, often accompanied by hashtags like #bangsamelayu (Malay race) and #13mei (May 13), were primarily created by seemingly Malay users using local languages and dialects.[5]

The spread of such content raised serious concerns among Malaysians, who feared that the hateful narratives could lead to real-world violence and threaten the country's fragile multi-ethnic harmony. The Malaysian government summoned ByteDance representatives to explain why such content was allowed on their platform. Following this meeting, TikTok removed thousands of offending posts and videos. However, despite the government's demands for a comprehensive crackdown, some problematic posts from the election period remained accessible on the platform[6] months after the election ended, highlighting the challenges in effectively moderating harmful content.

In a separate study conducted by Global Witness and the Cybersecurity for Democracy team at New York University in 2022,[7] researchers found that TikTok failed to catch 90 percent of ads featuring false and misleading messages about elections, while YouTube and Facebook identified and blocked most of them. The test involved submitting 10 ads in English and 10 in Spanish to the social media services using dummy accounts, without declaring the ads as political in nature or submitting to an identity verification process. Each ad, which included details like an incorrect election date or information designed to delegitimise the voting process, violated policies established by the respective platforms. TikTok's failure to reject these ads raises serious concerns about its ability to combat election-related disinformation.

Also, in June 2024, TikTok repeated a similar error by approving 16 advertisements targeted to Ireland containing[8] election disinformation ahead of the European parliamentary elections, as revealed by a Global Witness investigation. The adverts included false information encouraging people to vote online and by text, neither of which are permitted methods of voting in the upcoming elections, as well as false information about the voting age and incitement of force against immigrant voters. These findings raise questions about whether TikTok is breaching new EU rules that require platforms to mitigate election disinformation.

These case studies underscore the urgent need for social media companies to invest in more effective content moderation practices, particularly in non-English languages and during sensitive political events. The failure to do so can have severe consequences, including the erosion of democratic processes, the incitement of violence, and the undermining of social cohesion. As governments and civil society organisations continue to scrutinise the role of social media in shaping public discourse, it is imperative that these companies take decisive action to address the spread of hate speech, disinformation, and inflammatory content on their platforms.

Table 1: Some cases associated with social media platforms approving problematic contents in the past 8 years

Platform	Parent Company	Case
Facebook	Meta	• Hyper-partisan political ads and fake news generated more engagement on Facebook during the 2016 presidential election[9] • Meta's 2023 ad policy allowed posts that denied the legitimacy of the 2020 U.S. presidential election, enabling the spread of misinformation on its platform[10] • Meta approved AI manipulated political ads in India during 2024 Indian elections[11]
Instagram	Meta	• Meta approved ads on both Facebook and Instagram claiming 2020 Election was rigged.[12]
TikTok	Byte Dance	• TikTok approved 90% of 2022 US midterm election disinformation ads[13] • TikTok approved misleading election disinformation ads for publication in Ireland ahead of 2024 EU elections[14]
X	X	• X ran ads on #whitepower and other hate hashtags in June, 2024[15]
YouTube	Google	• 2021 YouTube continues to push dangerous videos to users susceptible to extremism, and white supremacy[16]

THE PERILS OF ENGAGEMENT-DRIVEN ALGORITHMS

At the heart of the problem lies the fundamental business model of social media companies. These platforms are designed to keep users scrolling, clicking, and engaging for as long as possible, as this allows them to serve more ads and generate more revenue. To achieve this goal, their algorithms are optimised to show users content that is most likely to capture their attention and elicit a strong emotional response. Unfortunately, research has shown that inflammatory, divisive, and sensationalistic content often drives the highest levels of engagement.[17] A 2018 study by researchers at MIT[18] found that false news stories spread six times faster on Twitter than true ones, and that lies were 70% more likely to be retweeted. Another study by researchers at New York University and Université Grenoble Alpes[19] found that false news received six times more likes, shares, and interactions on Facebook during the US 2020 election as compared to factual ones. These studies, among others, demonstrate that despite various mitigation efforts, misinformation and sensationalised content thrive on social media due to their ability to generate high user engagement.

Consequently, a perverse incentive is created for these algorithms to amplify hate speech, conspiracy theories, and extreme political views, since these are more likely to go viral and keep users hooked on the platform. This dynamic perpetuates a vicious cycle, where the most engaging content is promoted, regardless of its veracity or potential for harm, ultimately prioritising profits over the well-being of individuals and society as a whole.

Moreover, social media companies have become heavily reliant on political advertising[20] as a key revenue stream. During election campaigns, political parties and their supporters are willing to spend vast sums to promote their messages and target specific demographics. This has led to a perception that platforms like TikTok,[21] Facebook,[22] and X[23] are reluctant to crack down on misleading or inflammatory political ads, even when they violate their own policies. These companies often cite reasons such as freedom of expression,[24] mistakes in content moderation[25] or the inability to effectively monitor the vast amounts of content on their platforms.[26] But, the validity of these excuses is questionable, particularly given the substantial resources at their disposal.

Furthermore, the high cost of content moderation initiatives in non-major world languages including indigenous languages,[27] creates additional barriers to effective content regulation. This leaves marginalised communities particularly vulnerable to the spread of disinformation and hate speech. As a result of these inadequacies, a permissive environment has emerged where bad actors can spread disinformation and hate with relative impunity, knowing that social media companies will prioritise their ad dollars over the integrity of public discourse.

THE PATH FORWARD: STRONGER REGULATION AND GLOBAL COOPERATION

It is clear that social media companies' attempts at self-regulation have been woefully inadequate. Voluntary measures and public relations campaigns have failed to address the systemic issues that allow hate and disinformation to flourish on these platforms. Governments and international organisations must step in to hold these companies accountable and enforce meaningful standards for content moderation.

One potential approach is the imposition of substantial fines for repeated failures to enforce content policies. If social media giants and to face significant financial penalties each time they allowed hate speech or disinformation to spread, they would be far more motivated to invest in robust moderation systems and human oversight. Another strategy could involve mandating a minimum level of investment in content moderation, especially for non-dominant, non-English languages, where harmful content frequently slips through the cracks. Regulators could require that a specific percentage of these companies' revenue be dedicated to hiring and training moderators with the linguistic and cultural knowledge needed to identify and remove problematic posts.

Regular third-party audits of content moderation systems and ad approval processes could also play a crucial role in promoting transparency and accountability. Independent auditors could assess the effectiveness of these systems, identify areas for improvement, and publicly report on their findings. This would provide much-needed visibility into the inner workings of these platforms and put pressure on companies to address any shortcomings. Additionally, policymakers may need to reexamine the legal frameworks that currently shield social media companies from liability for user-generated content. If it can be demonstrated that their algorithms are systematically amplifying hate speech and disinformation, they may need to bear greater responsibility for the harms that result.

However, given the global nature of social media, a piecemeal approach by individual countries is unlikely to be sufficient. To truly rein in the power of these platforms, we need coordinated action at the regional and international levels will be needed. Countries facing similar challenges with online hate and disinformation should come together in demanding change from social media companies. For example, member states of the Association of Southeast Asian Nations (ASEAN) could greatly benefit from collaborating to develop a shared set of standards and regulations for content moderation. While achieving broad consensus on content moderation standards across ASEAN member states may be difficult due to diverse political systems and interests, there are still opportunities for cooperation on specific issues of shared concern.

ASEAN governments could focus on areas where there is greater alignment, such as combating online exploitation, protecting minors, and countering violent extremism. By pooling resources and expertise, member states can develop targeted initiatives to tackle these pressing issues more effectively. Additionally, regional cooperation should prioritise capacity building and knowledge sharing to help governments navigate the complexities of online content regulation. This could include joint research projects, training programmes for policymakers and regulators, and platforms for ongoing dialogue and coordination. By fostering a shared understanding of the challenges and best practices in content moderation, ASEAN member states can work towards more informed and effective policymaking.

While top-down regulation may be challenging given the political dynamics within ASEAN, member states can still advocate for greater transparency and accountability from social media companies. This could involve pushing for more investment in local content moderation teams, clearer timelines for removing flagged content, and increased transparency around algorithms and data practices.

To complement regional efforts, ASEAN could also engage with multilateral forums such as the United Nations, the G20, or the OECD to develop global norms and guidelines on social media governance. By sharing their unique perspectives and experiences, ASEAN member states can help shape the international discourse on these critical issues.

Ultimately, while the path to effective content moderation in Southeast Asia may be complex, a flexible and collaborative approach that respects the diverse contexts of member states offers the best chance of making meaningful progress. By focusing on areas of common ground and working towards incremental improvements, ASEAN governments can help create a safer and more responsible online environment for their citizens.

SAFEGUARDING DEMOCRACY IN THE DIGITAL AGE

The unchecked power of social media companies poses a danger to the health of democracies. The case studies from India and Malaysia demonstrate how the failure to effectively moderate online content can fuel real-world violence, deepen societal divisions, and erode public trust in democratic institutions, especially in the age of rapidly developing communication technology and the increasing penetration of AI into our daily lives.

We have seen the deadly impact of social media-fuelled violence[28] in places like Myanmar[29] and Sri Lanka,[30] the erosion of public trust in democratic institutions as a result of targeted disinformation campaigns and the rise of polarisation and extremism as algorithms feed users' increasingly radical content. It is time for policymakers, civil society, and the public to demand accountability from social media giants and take decisive action to address the harms they enable. This will require a multi-faceted approach, combining stronger regulation, increased investment in content moderation, regular audits and transparency measures, and legal reforms to clarify the responsibilities of these companies. Close cooperation and coordination among countries at regional and global levels are crucial to ensure consistent, effective, and enforceable standards and rules governing social media. The stakes could not be higher. If we fail to act, we risk allowing social media companies to continue to be used as weapons against the very values and institutions that underpin our free and open societies.

The power of social media companies is too great to be left unchecked. It is time to work together to build a digital future that promotes transparency, accountability, and the responsible exchange of ideas, while safeguarding against the spread of hate, lies, and division. Only then can we fully realise the potential of these technologies to inform, connect, and empower us, rather than divide and mislead us.

ENDNOTES

[1] Hannah Ellis-Pettersen, "Revealed: Meta approved political ads in India that incited violence." *The Guardian.* 20 May 2024, https://www.theguardian.com/world/article/2024/may/20/revealed-meta-approved-political-ads-in-india-that-incited-violence

[2] Witness Org. 3 November 2022, https://www.globalwitness.org/en/campaigns/digital-threats/open-door-hate-meta-approves-ads-containing-far-right-hate-speech-norwegian/

[3] Francessa Visser and Pri Bengani, "Facebook failed to block thousands of political ads peddling false information." *The Bureau of Investigative Journalism.* 05 May 2024, https://www.thebureauinvestigates.com/stories/2024-06-05/facebook-failed-to-block-thousands-of-political-ads-peddling-false-information/

[4] Rozanna Latif and Mei Mei Chu, "TikTok on high alert in Malaysia as tensions rise over election wrangle." *Reuters.* 23 November 2022. https://www.reuters.com/world/asia-pacific/tiktok-high-alert-malaysia-tensions-rise-over-election-wrangle-2022-11-23/

[5] Nuurrianti Jalli, "Disinformation and Hate Speech: Ethnoreligious Rhetoric on TikTok during Malaysia's 15th General Election (GE15) 2022." Konferensi Ilmu Sosial dan Ilmu Politik KISIP 2024. https://shareok.org/handle/11244/340152

[6] Nuurrianti Jalli "How TikTok became a breeding ground for hate speech in the latest Malaysia general election." *The Conversation.* 23 March 2023. https://theconversation.com/how-tiktok-became-a-breeding-ground-for-hate-speech-in-the-latest-malaysia-general-election-200542

[7] Global Witness. "TikTok and Facebook fail to detect election disinformation in the US, while YouTube succeeds." *Global Witness Org.* 21 October 2022. https://www.globalwitness.org/en/campaigns/digital-threats/tiktok-and-facebook-fail-detect-election-disinformation-us-while-youtube-succeeds/

[8] Global Witness. "TikTok approves misleading election disinformation ads for publication in Ireland ahead of EU elections." *Global Witness Org*. 4 June 2024. https://www.globalwitness.org/en/press-releases/tiktok-approves-misleading-election-disinformation-ads-publication-ireland-ahead-eu-elections/

[9] Anita Balakrishnan. "More Facebook users engaged with top fake election news than most popular real reporting, report says." *CNBC*. 16 November 2016. https://www.cnbc.com/2016/11/16/more-facebook-users-engaged-with-top-fake-election-news-than-most-popular-real-reporting-report-says.html

[10] Max Zahn. "Meta ad policy allowing 2020 election denial followed warning of political backlash, sources say." *ABC News.* 14 December 2023. https://abcnews.go.com/US/meta-ad-policy-allowing-2020-election-denial-warning/story?id=105437974

[11] Ekō. "Meta fails to stop violent and inflammatory AI-generated ads targeting Indian voters." *Ekō.org.* 16 May 2024. https://www.eko.org/media/meta-fails-to-stop-violent-and-inflammatory-ai-generated-ads-targeting-indian-voters/

[12] Salvador Rodriguez. "Meta allows ads claiming rigged 2020 election on Facebook, Instagram." *The Wall Street Journal.* 15 November 2023. https://www.wsj.com/tech/meta-allows-ads-claiming-rigged-2020-election-on-facebook-instagram-309b678d

[13] Global Witness. "TikTok and Facebook fail to detect election disinformation in the US, while YouTube succeeds." *Global Witness Org.* 21 October 2022. https://www.globalwitness.org/en/campaigns/digital-threats/tiktok-and-facebook-fail-detect-election-disinformation-us-while-youtube-succeeds/

[14] Global Witness. "TikTok approves misleading election disinformation ads for publication in Ireland ahead of EU elections." *Global Witness Org*. 4 June 2024. https://www.globalwitness.org/en/press-releases/tiktok-approves-misleading-election-disinformation-ads-publication-ireland-ahead-eu-elections/

[15] David Ingram. "Elon Musk's X app ran ads on #whitepower and other hateful hashtags." *NBC News.* 6 June 2024. https://www.nbcnews.com/tech/social-media/elon-musk-x-twitter-antisemitism-hashtags-trending-hate-rcna151945

[16] Annie Y Chen, et. al. "Exposure to alternative and extremist content on YouTube." *Anti-Defamation League (ADL).* 5 February 2021. https://www.adl.org/resources/report/exposure-alternative-extremist-content-youtube

[17] Orion Arthur Yoesle. *"I am human being, dammit!" – partisan media, sociopolitical identity, emotional arousal, and incivility seeking behavior."* Washington State University. May 2023. https://search.proquest.com/openview/8ebb6955c155e2d214f6c94a6b61a13e/1.pdf?pq-origsite=gscholar&cbl=18750&diss=y

[18] Peter Dizikes. "Study: On Twitter, false news travels faster than true stories." *MIT News.* 8 March 2018. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[19] Elizabeth Dwoskin. "Misinformation on Facebook got six times more clicks than factual news during the 2020 election, study says." *The Washington Post.* 4 September 2021. https://www.washingtonpost.com/technology/2021/09/03/facebook-misinformation-nyu-study/

[20] Alexa Raad. "It's Time for Congress to Curb Political Ads on Social Media." *Centre for International Governance Innovation*. 9 August 2023. https://www.cigionline.org/articles/political-ads-on-social-media-have-sown-division-confusion-suppression-congress-needs-to-act-now/

[21] Nuurrianti Jalli and Ika Idris. "Misinformation, bias: time ticking for TikTok to review policies." *The Vibes.* 29 June 2023. https://www.thevibes.com/articles/opinion/95518/misinformation-bias-time-ticking-for-tiktok-to-review-policies-nuurrianti-jalli-ika-idris

[22] Jeff Horwitz and Deepa Seetharaman. "Facebook Executives Shut Down Efforts to Make the Site Less Divisive." *The Wall Street Journal.* 26 May 2020. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

[23] Kelvin Chan. "IBM, EU and Lionsgate pull ads from X over concerns of antisemitic messages." *PBS.* 17 November 2023. https://www.pbs.org/newshour/economy/ibm-pulls-advertising-from-x-after-report-shows-ads-ran-next-to-antisemitic-messages

[24] Alexis Madrigal. "The 'Platform' Excuse Is Dying." *The Atlantic.* 11 June 2019. https://www.theatlantic.com/technology/archive/2019/06/facebook-and-youtubes-platform-excuse-dying/591466/

[25] Charlotte Jee. “Facebook needs 30,000 of its own content moderators, says a new report.” *MIT Technology Review.* 8 June 2020. https://www.technologyreview.com/2020/06/08/1002894/facebook-needs-30000-of-its-own-content-moderators-says-a-new-report/#:~:text=Errors%20are%20rife%2C%20despite%20the,kept%20up%20or%20vice%20versa.

[26] Rem Darbinyan. “The growing role of AI in content moderation.” *Forbes*. 14 June 2022. https://www.forbes.com/sites/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation/

[27] Global Witness. “How Big Tech platforms are neglecting their non-English language users”. *Global Witness Org*. 30 November 2023. https://www.globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/

[28] Zachary Laub. “Hate Speech on Social Media: Global Comparisons.” *Council on Foreign Relations.* 7 June 2019. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

[29] Amnesty International. “Myanmar: Facebook’s systems promoted violence against Rohingya; Meta owes reparations – new report.” *Amnesty International.* 29 September 2022. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/#:~:text=Facebook%20owner%20Meta%27s%20dangerous%20algorithms,a%20new%20report%20published%20today.

[30] Shihar Aneez and Ranga Sirilal. “Sri Lanka to lift social media ban: minister.” *Reuters.* 15 March 2018. https://www.reuters.com/article/us-sri-lanka-clashes-socialmedia/sri-lanka-to-lift-social-media-ban-minister-idUSKCN1GP2LO/

ISEAS Perspective is published electronically by: **ISEAS - Yusof Ishak Institute** 30 Heng Mui Keng Terrace Singapore 119614 Main Tel: (65) 6778 0955 Main Fax: (65) 6778 1735 Get Involved with ISEAS. Please click here: https://www.iseas.edu.sg/support/get-involved-with-iseas/	ISEAS - Yusof Ishak Institute accepts no responsibility for facts presented and views expressed. Responsibility rests exclusively with the individual author or authors. No part of this publication may be reproduced in any form without permission. © Copyright is held by the author or authors of each article.	Editorial Chairman: Choi Shing Kwok Editorial Advisor: Tan Chin Tiong Editorial Committee: Terence Chong, Cassey Lee, Norshahril Saat, and Hoang Thi Ha Managing Editor: Ooi Kee Beng Editors: William Choong, Lee Poh Onn, Lee Sue-Ann, and Ng Kah Meng Comments are welcome and may be sent to the author(s).