PERSPECTIVE

RESEARCHERS AT ISEAS – YUSOF ISHAK INSTITUTE ANALYSE CURRENT EVENTS

Singapore | 4 August 2025

From Principles to Protocols: Embedding Partnerships into Content Moderation Technologies Against Mis/Disinformation

Beltsazar Krisetya*

A man watches a video of misinformation on the social media platform TikTok on his mobile phone in Hanoi on 6 October 2023. (Photo by Nhac NGUYEN / AFP).

* Beltsazar Krisetya [1] is Visiting Fellow at the Media, Technology, and Society (MTS) Programme, ISEAS-Yusof Ishak Institute; PhD Student in Science, Technology, Engineering, and Public Policy (STEaPP) at University College London (UCL); and Researcher (on study leave) at the Centre for Strategic and International Studies (CSIS).

EXECUTIVE SUMMARY

- Despite growing calls for collaboration, the technological core of content moderation remains largely a black box. While multi-stakeholder partnerships are increasingly invoked in regional policy discourse, external oversight or shared governance of the underlying moderation technologies remain limited.

- Content moderation technologies differ in function and ownership. Tools used to detect clearly illegal content (e.g. terrorist material, child exploitation) are slightly more established and collaborative, while technologies for nuanced borderline content (e.g. misinformation) remain experimental and proprietary.

- While technological control is concentrated, partnerships are fragmented. Even where external partnerships exist (e.g. for fact-checking or hash-sharing), they are often optional, ad hoc, or supplementary. Platform incentives, regulatory gaps, and technical complexity collectively reinforce a system where meaningful input from governments or civil society remains limited.

- Even if distant, the idea of "partnership by design" is worth further exploration. Stakeholders need to entertain the feasibility of embedding multi-stakeholder input directly into the architecture of content moderation systems, especially for emerging tools such as LLMs and NLP. This requires governments to articulate clear standards and platforms for such purposes from the outset.

INTRODUCTION

In addressing false and harmful online content, "multi-stakeholder partnership" has become a catch-all phrase to describe any kind of collaboration between actors. At the normative level, the 2023 ASEAN Guideline on Management of Government Information in Combating Fake News and Disinformation in the Media[2] suggests the "penta-helix" approach. Coordinated by Southeast Asian governments, it invites partnerships between business, media, civil society, academia, and governments to debunk and amplify counter-narratives against disinformation.

Institutionally, governments have established formal multi-stakeholder platforms to shape online content. For example, Malaysia's Communications and Multimedia Content Forum (CMCF),[3] a self-regulatory body, draws multiple actors to co-draft and enforce a Content Code and manage public complaints. In the Philippines, the government-led Inter-Agency Council Against Child Pornography (IACACP) [4] coordinates policy rollout, capacity-building, advocacy, and access blocking to curb online child sexual exploitation.

Partnerships also occur in platform-led initiatives, which might include Priority Flagger (YouTube), Trusted Partner (Meta), or Safety Partners (TikTok) programmes, essentially providing a highway for partner organisations to report problematic content to platforms.

However, while partnerships have begun to materialise at various levels, the technological layer of content moderation remains largely proprietary, i.e. controlled and owned by tech platforms. This article examines the elements of technologies that make up content moderation (particularly against borderline content) and assesses whether each aspect can be grounds for further collaboration or, conversely, whether they are a site of power struggle.
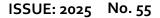
THE COMPLEX SYSTEM OF CONTENT MODERATION

Content moderation, understood as the assessment of user-generated content (UGC) to determine appropriateness, [5] comprises a complex and interconnected set of standards, practices, and technological design. Each of those elements eventually affects the moderation outcomes directly and indirectly.

Zooming in on the technological aspect in particular, the policy discourses in Southeast Asia tend to view the social media "algorithm" as a single entity rather than recognising it as a collection of diverse automated systems. Non-algorithm elements also risk being overlooked, not least because policymakers tend to focus more on the *purpose* and *outcomes* of such algorithms, instead of the algorithm's technical possibilities, development life cycle, and ownership/control.

The output-driven approach can indeed be helpful for devising normative and strategic guidelines. However, this might come at the expense of tailored, surgical interventions to address specific technical harms, risks, or biases, that may widen the already sizeable gap between policy intent and implementation.

To understand the technological nuts and bolts relevant for better multi-stakeholder policies, the table below presents a non-exhaustive list of the components of the technical system that makes up content moderation. This analysis will use the umbrella term of "technology" to discuss techniques, platform-deployed tools/system, proprietary tools, or hybrid enforcement frameworks that make up content moderation.

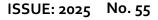| Type	Technology	Description
Foundational AI and data processing techniques	*Machine Learning (ML)*	Algorithms trained to identify patterns and flag content without hard rules.[6]
	Hashing (Cryptographic)	Detects exact matches of banned files using unique digital fingerprints.[7]
Deep Learning (DL)	*Natural Language Processing (NLP)*	Analyses human language in text for harmful or policy-violating content.[8]
	Large Language Models (LLMs)	AI systems that understand and generate human-like text across languages.[9]
	Video Frame Analysis	AI reviews video images to detect nudity, violence, and other visual harms.[10]
	Audio Analysis	Scans audio for forbidden content or matches to copyright/terror databases.[11]
	Deepfake Detection	Identifies AI-generated fake media like videos or voice clones.[12]
Pattern and Behaviour Analysis	*Coordinated Inauthentic Behaviour (CIB) Detection*	Flags networks of fake or deceptive accounts acting in coordination.[13]
	Keyword Filtering	Blocks content with pre-set banned terms.[14]
	Hashing (Perceptual)	Identifies visually similar content even if slightly altered.[15]
Human-Centred Frameworks	*Human-in-the-Loop (HITL)*	Combines AI speed with human judgment and correction.[16]
	Community Moderation	Users flag content or add public context (e.g. Community Notes).[17]

There are, of course, content moderation elements not primarily driven by technology but by processes or people. For example, there are the Business Process Outsourcing (BPOs) model for content moderators, [18] the aforementioned trusted partner programme, and also independent/third-party fact-checkers.

The moderation process also refers to several, sometimes competing, written guidelines in national legislations, platform community standards, voluntary codes of conduct,[19] or global norms like the Santa Clara Principles.[20] Moderation outcomes are also not immune to government-mandated takedown or Oversight Board decisions [21] (in the case of Meta platforms). These non-technological elements might be endless; however, the bottom line is that the technologies do not exist in a silo but are influenced by (and in turn influence) other elements of content moderation.

Additionally, the list only includes established examples and deliberately omits experimental moderation concepts, albeit some of them might be promising. For example, Francis Fukuyama[22] proposes the concept of "middleware", a third-party software positioned between users and tech giants, giving users power to choose how the content is curated rather than just accepting platforms' default algorithms. This idea has materialised in newer/smaller platforms such as BlueSky,[23] but requires tremendous buy-in to be implemented in more mainstream platforms.

This article focuses the technological aspects of content moderation. The technologies involved in content moderation are categorised according to two criteria: whether the technology is

reliable for borderline/nuanced content, and whether the existing governance model is proprietary or already has some form of collaboration with external actors. These are reflected in the 2x2 matrix below:[24]
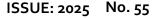
	Only effective for clearly harmful and illegal content	*Also promising for borderline/context-dependent content*
Requires external input/partnership (shared governance or input)	*Quadrant II* • Cryptographic hashing • Perceptual hashing • Deepfake detection	*Quadrant I* • Community moderation/notes • Human-in-the-Loop (HITL) with fact-checkers
Platform-controlled (proprietary/closed system)	*Quadrant III* • Machine Learning (ML) • Keyword filtering • Audio fingerprinting • Video frame analysis	*Quadrant IV* • Natural Language Processing (NLP) • Large Language Models (LLMs) • Internal Human-in-the-Loop (HITL) • Manual moderation • Detection system for Coordinated Inauthentic Behaviour (CIB)

Ideally, there should be more technologies in Quadrant I, where multi-stakeholder partnerships in technology can address the more complex borderline and context-dependent content. The technologies listed in Quadrant II represent the low-hanging fruit for multi-stakeholder cooperation, since technologies like ContentID or hashing are only reliable in addressing clearly defined and easily identified illegal content. Thus, such partnerships do not entail high-level decision-making or complex problem solving. Quadrant III lists the proprietary technologies in complete or almost complete control of tech platforms that have been reliably deployed for easily identified illegal contents, but not so much for borderline content. Finally, Quadrant IV lists the technologies that have potential to be effectively deployed for borderline content but are still largely controlled by tech platforms.

INHERENT LIMITATIONS

There are clear observations from the matrix. The most obvious one is that despite the emerging multi-stakeholder partnerships in fact-checking programmes and shaping regulations, the technological layer of content moderation is still largely privately controlled by tech platforms (Quadrants III and IV). After all, it is the platforms that have put in substantial infrastructure investment and provide maintenance of those systems.

This arrangement, combined with external factors such as under/overregulation and the inherent complexities of content moderation, further incentivises the platforms to maintain

proprietary systems rather than engage in broader collective efforts. As a result, there is currently little room for Southeast Asian countries and regional organisations to exert their influence on technologies owned and implemented by US or China-based social media platforms.

The second observation is that the technologies that address easily identified illegal contents (Quadrants II and III) are relatively more established than the emerging/experimental technologies for borderline content. Established technologies generally enable a more defined partnership. For example, moderation against terrorism and violent extremism content has long benefited from the hash-sharing database[25] run by the Global Internet Forum to Counter Terrorism (GIFCT), a non-profit consortium founded by Meta, Microsoft, YouTube, and then-Twitter.

The GIFCT initiative converts known terrorist or violent extremist content into unique representations known as *hashes*. Member companies can then quickly block and remove content matching those hashes without needing to share user data or original files. Misinformation and disinformation, however, often lack a clear and consistent digital signature and consistently shapeshifts—whether through text alteration, image manipulation, or misleading narratives that evade standard detection algorithms. These "lawful but awful" contents exploit not only legal limitations,[26] but also a technological one.

Third, technology that has the potential to address borderline contents requires substantial human participation in its life cycle. As categorised in Quadrant I, human judgement is a key component in both community notes and Human-in-the-Loop (HITL) with fact-checkers. Community notes harness the collective evaluation of social media users to provide additional context to the original posts.[27] HITL with fact-checkers combines platforms' internal HITL system with external input with third-party fact-checkers.

Such partnerships, however, are currently limited and optional. HITL can easily be achieved with internal moderators within social media platforms instead of third-party fact checkers (see Quadrant IV). Community notes, which recently re-emerged after being adapted by Meta platforms (in the US market only), are largely limited to ex-post interventions – i.e., providing additional context after mis/disinformation had already been posted and deemed too slow in reducing engagement with mis/disinformation during the early stage of dissemination.[28] This makes the existing interventions in Quadrant I optional and supplementary, as opposed to being prioritised as an essential tool.

Finally, underneath the technologies managed by social media platforms, there is still a matter of government internet-filtering technologies. To some authoritarian regimes, there is no need to moderate content if no objectionable content can be uploaded in the first place. Southeast Asia is no stranger to instances of governments engaging in internet filtering, whether generally through DNS/IP blocking or surgically through Deep Packet Inspection (DPI).[29] These instruments are traditionally employed to address illegal content but have been increasingly used for borderline content.[30]

Social media platforms may have technological prowess and trade secrets, but governments still hold the internet's "kill switch". Whereas undersea cables[31] can be a site of power struggle between governments, content moderation becomes a similar arena for contestation between governments and corporations.

GOING FORWARD

Effective regulatory efforts in the region will only be implementable to the extent that the technological limitations allow. The technologies in Quadrant IV represent the last frontier for multi-stakeholder partnership due to their emergent nature, proprietary ownership, and difficulty in achieving consensus on moderating borderline contents.

The policy implications are two-fold. The first relates to the importance of differentiating between regulation for clear-cut illegal/harmful content versus borderline content. Most ASEAN member states still treat both kinds of content using the same legal provisions, with the exception of Singapore with the Online Safety (Miscellaneous Amendments) Act for harmful content and the Protection from Online Falsehoods and Manipulation Act (POFMA) for borderline content. While POFMA is not without its weaknesses, it still underlines the importance of treating illegal and borderline contents separately, and under two separate legislative frameworks and processes.

Second, these is a need to specify technical regulations for each technological element of content moderation operated by tech platforms. This is trickier to articulate, let alone implement. However, metaphorically, as much as the proprietary algorithms constitute the company's "secret sauce", such sauces should not be exempt from food safety standards.

Elsewhere in the digital policy landscape, there has been an emergence of "by design" approaches. Fundamentally, "by design" is a proactive act to systematically embed a certain principle in the foundation of a certain technology and its life cycle, rather than adding them as an afterthought. There is a "secure-by-design" approach in the Philippines National Cyber Security Plan 2023 – 2028,[32] "safety by design" in the Indonesia's Online Child Protection Regulation,[33] and "privacy by design" guidelines set to be developed by Malaysia to supplement the revised Personal Data Protection Act.[34]

Following a similar vein, it might be beneficial for policymakers to explore the notion of "partnership by design" for content moderation's technical architecture. This approach can benefit borderline content (such as misinformation and disinformation) by expanding the room in which non-corporate stakeholders can provide direct input to content moderation technologies.

The governments' part of the job is to pronounce what exactly "partnership by design" means and in which technologies they want the partnership to happen. For instance, with the emergence of explainable AI (XAI) in policy discourse,[35] government could specify that for content moderated by technologies in Quadrant IV (e.g., NLP, LLMs, internal HITL), platforms must be able to demonstrate the specific linguistic cues, contextual elements, or data correlations that led to an automated moderation decision. This has become more imperative given the recent trend of platforms to downsize its Trust and Safety (T&S) team in favour of AI moderation.[36]

The platforms' part of the job is to lead the partnership to ensure that the standards are implementable and operable across platform. Learning from the best practices of GIFCT and the Coalition for Content Provenance and Authenticity (C2PA), buy-ins from tech platforms are easier if the initiatives are platform-led, rather than government-led.

Admittedly, this proposition is intended to provoke, rather than prescribe. After all, there is an inherent impasse: social media platforms are understandably reluctant to improve transparency without clear regulatory protection, while regulators struggle to craft and operationalise policies without access to the technical insights and capabilities held by the platforms. Most Southeast Asian countries are also "tech-takers", rather than "tech-makers", which significantly reduces their leverage and ability to effectively regulate tech platforms in the overall scheme of things.

Nonetheless, the central point stands: normative agreements and ad hoc programmes between stakeholders alone are not enough. They need to be translated—not just into binding regulations—but also into the binary codes that make up the technical system of content moderation. Multi-stakeholder cooperation needs *operationalisation*, as much as it needs *institutionalisation*.

ENDNOTES

[1] The author declares no conflict of interest, and wants to thank Lee Sue-Ann, Maria Monica Wihardja, Roser Pujadas, and Diyi Liu for feedback on the initial draft.

[2] Ministry of Communication and Informatics of the Republic of Indonesia, *ASEAN Guideline on Management of Government Information in Combating Fake News and Disinformation in The Media* (ASEAN, n.d.), accessed 30 June 2025, https://asean.org/book/asean-guideline-on-management-of-government-information-in-combating-fake-news-and-disinformation-in-the-media/.

[3] Communications and Multimedia Content Forum, 'Content Forum', Communications and Multimedia Content Forum, accessed 22 July 2025, https://contentforum.my/.

[4] IACACP, *The Inter-Agency Council against Child Pornography*, n.d., accessed 22 July 2025, https://www.iacacp.gov.ph/.

[5] Pauline Leong, 'Content Moderation of Social Media in Southeast Asia: Contestations and Control', Fulcrum, 27 July 2022, https://fulcrum.sg/content-moderation-of-social-media-in-southeast-asia-contestations-and-control/.

[6] Juan Felipe Gomez et al., 'Algorithmic Arbitrariness in Content Moderation', *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 3 June 2024, 2234–53, https://doi.org/10.1145/3630106.3659036.

[7] CometChat, 'What Is Hash Matching? How It Helps in CSAM Detection?', CometChat, accessed 9 July 2025, https://www.cometchat.com/blog/what-is-hash-matching.

[8] CloudFlare, 'What Is NLP in AI?', Natural Language Processing, accessed 9 July 2025, https://www.cloudflare.com/learning/ai/natural-language-processing-nlp/.

[9] WhyLabs, 'Content Moderation with Large Language Models (LLMs)', WhyLabs, accessed 9 July 2025, https://whylabs.ai/learning-center/llm-use-cases/content-moderation-with-large-language-models-llms.

[10] Jagreet Kaur, 'AI-Powered Video Analytics for Real-Time Content Moderation', XenonStack, 2025, https://www.xenonstack.com/blog/ai-powered-video-analytics.

[11] Stream, 'Audio Moderation - What Is It and How Does It Work?', Stream, accessed 9 July 2025, https://getstream.io/glossary/audio-moderation/.

[12] Lauren Hendrickson, 'Deepfake Detection: How to Spot and Prevent Synthetic Media', *Identity*, 23 June 2025, https://www.identity.com/deepfake-detection-how-to-spot-and-prevent-synthetic-media/.

[13] Meta, 'Inauthentic Behaviour', Transparency Centre, 2025, https://transparency.meta.com/en-gb/policies/community-standards/inauthentic-behavior/.

[14] Faruk Alpay, 'Bypassing Content Moderation Filters: Techniques, Challenges, and Implications', Lightcap AI, 13 May 2025, https://lightcapai.medium.com/bypassing-content-moderation-filters-techniques-challenges-and-implications-4d329f43a6c1.
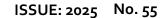
[15] Sean McKeown and William J Buchanan, *Hamming Distributions of Popular Perceptual Hashing Techniques*, version 1, arXiv, 2022, https://doi.org/10.48550/ARXIV.2212.08035.

[16] TELUS Digital, 'What Is Human-in-the-Loop?', accessed 9 July 2025, https://www.telusdigital.com/glossary/human-in-the-loop.

[17] Khoros, '11 Best Practices for Online Community Moderation', 2025, https://khoros.com/blog/community-moderation.

[18] Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press, 2019).

[19] Stephan Mündges and Kirsty Park, 'But Did They Really? Platforms' Compliance with the Code of Practice on Disinformation in Review', *Internet Policy Review* 13, no. 3 (2024), https://doi.org/10.14763/2024.3.1786.

[20] 'Santa Clara Principles on Transparency and Accountability in Content Moderation', Santa Clara Principles, accessed 9 July 2025, https://santaclaraprinciples.org/images/santa-clara-OG.png.

[21] Evelyn Douek, 'The Meta Oversight Board and the Empty Promise of Legitimacy', SSRN Scholarly Paper no. 4565180 (Social Science Research Network, 7 September 2023), https://doi.org/10.2139/ssrn.4565180.

[22] Francis Fukuyama, 'Radical Proposal: Middleware Could Give Consumers Choices Over What They See Online', Stanford HAI, 2021, https://hai.stanford.edu/news/radical-proposal-middleware-could-give-consumers-choices-over-what-they-see-online.

[23] Luke Hogg et al., 'Shaping the Future of Social Media with Middleware', arXiv:2412.10283, preprint, arXiv, 13 December 2024, https://doi.org/10.48550/arXiv.2412.10283.

[24] Sahana Udupa et al., 'Ethical Scaling for Content Moderation: Extreme Speech and the (in)Significance of Artificial Intelligence', *Big Data & Society* 10, no. 1 (2023), https://doi.org/10.1177/20539517231172424; Arkaitz Zubiaga, 'Natural Language Processing in the Era of Large Language Models', *Frontiers in Artificial Intelligence* 6 (January 2024), https://doi.org/10.3389/frai.2023.1350306; Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, 2018); Valentine Crosset and Benoît Dupont, 'Cognitive Assemblages: The Entangled Nature of Algorithmic Content Moderation', *Big Data & Society* 9, no. 2 (2022): 20539517221143361, https://doi.org/10.1177/20539517221143361; Todor Markov et al., 'A Holistic Approach to Undesired Content Detection in the Real World', *Proceedings of the AAAI Conference on Artificial Intelligence* 37, no. 12 (2023): 15009–18, https://doi.org/10.1609/aaai.v37i12.26752; Ashley A. Mattheis and Ashton Kingdon, 'Moderating Manipulation: Demystifying Extremist Tactics for Gaming the (Regulatory) System', *Policy & Internet* 15, no. 4 (2023): 478–97, https://doi.org/10.1002/poi3.381.

[25] Global Internet Forum to Counter Terrorism, 'GIFCT's Hash-Sharing Database', GIFCT, accessed 9 July 2025, https://gifct.org/hsdb/.

[26] Simon Chesterman, 'Lawful but Awful: Evolving Legislative Responses to Address Online Misinformation, Disinformation, and Mal-Information in the Age of Generative AI', *The American Journal of Comparative Law*, ahead of print, Oxford University Press (OUP), 7 July 2025, https://doi.org/10.1093/ajcl/avaf020.

[27] Jonathan Warden, 'Understanding Community Notes and Bridging-Based Ranking', 2024, https://jonathanwarden.com/understanding-community-notes/.

[28] Yuwei Chuai et al., 'Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?', *Proceedings of the ACM on Human-Computer Interaction* 8, no. CSCW2 (2024): 1–52, https://doi.org/10.1145/3686967.

[29] Doug Bernard, 'Thailand Set to Build China-like Internet Firewall', Voice of America, 28 September 2015, https://www.voanews.com/a/thailand-set-to-build-china-like-internet-firewall/2982650.html.

[30] Surachanee "Hammerli" Sriyai, *How Means for Digital Repression in Southeast Asia Have Unfolded in Recent Times*, 2024, no. No. 65 (2024), https://www.iseas.edu.sg/articles-commentaries/iseas-perspective/2024-65-how-means-for-digital-repression-in-southeast-asia-have-unfolded-in-recent-times-by-surachanee-hammerli-sriyai/.

[31] Elina Noor, 'Entangled: Southeast Asia and the Geopolitics of Undersea Cables', Center for Indo-Pacific Affairs, accessed 9 July 2025, https://manoa.hawaii.edu/indopacificaffairs/article/entangled-southeast-asia-and-the-geopolitics-of-undersea-cables/.

[32] Department of Information and Communications Technology, 'National Cyber Security Plan 2023 - 2028', Department of Information and Communications Technology, accessed 9 July 2025, https://dict.gov.ph/national-cyber-security-plan.

[33] JDIH BPK, 'Tata Kelola Penyelenggaraan Sistem Elektronik Dalam Pelindungan Anak', Database Peraturan, 2025, http://peraturan.bpk.go.id/Details/316698/pp-no-17-tahun-2025.

[34] Personal Data Protection Commissioner, 'Announcement on the Public Consultation Paper (PCP) on Personal Data Protection 2025 • Personal Data Protection', Personal Data Protection Commissioner, 2025, https://www.pdp.gov.my/ppdpv1/en/announcement-on-the-public-consultation-paper-pcp-on-personal-data-protection-2025/.

[35] 'What Explainable AI Is, Why It Matters and How We Can Achieve It - OECD.AI', accessed 22 July 2025, https://oecd.ai/en/wonk/what-explainable-ai-is-why-it-matters-and-how-we-can-achieve-it.

[36] Rachel Elizabeth Moran et al., 'The End of Trust and Safety?: Examining the Future of Content Moderation and Upheavals in Professional Online Safety Efforts', *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ACM, 26 April 2025, 1–14, https://doi.org/10.1145/3706598.3713662.