Safer Internet Lab
Led by CSIS in partnership with Google

Centre for Strategic and International Studies

Google

Snapshot

The Global Context of Social Media Governance

Beltsazar A. Krisetya | Medelina K. Hendytio | Debora Irene Christine | Sherly Haristya | Sekar Arum Jannah

Editor: Irene Poetranto

SAIL Snapshots is a peer-reviewed article series featuring key research findings from Safer Internet Lab researchers and its associates. The views presented are exclusively those of the authors and do not reflect the official stance of SAIL, CSIS, Google, or any other affiliated organisation.

The Global Context of Social Media Governance

Beltsazar A. Krisetya[1] Medelina K. Hendytio[2], Debora Irene Christine[3], Sherly Haristya[4], Sekar Arum Jannah[5]

Editor: Irene Poetranto[6]

There are at least three prevailing models of social media governance against information disorder: state regulation through national laws, voluntary self-regulation by platforms themselves, and co-regulation through collaborative initiatives involving platforms and other stakeholders.[7]

The State-Driven Regulatory Approach

Governments worldwide are responding to the rise of online misinformation and disinformation by introducing new laws and regulations. These emerging legal frameworks have taken varied approaches on several factors, such as the scope of regulations and the responsibilities of social media platforms.

Firstly, the scope of regulated content differs significantly across laws. Germany's Network Enforcement Act, known as NetzDG, focuses specifically on online hate speech and disinformation. NetzDG's narrow scope can facilitate oversight and enforcement. However, it relies heavily on the German criminal code, which does not fully address new disinformation tactics like micro-targeting and computational propaganda that exploit social media algorithms.[8]

In contrast, the European Union's Digital Services Act (DSA) aims to regulate a broader range of prohibited and illegal online content. The DSA proportionately assigns content moderation duties to platforms based on their scale and impact, such as the number of users. The obligation to moderate content is balanced with protections for users' free expression rights.[9] However, the DSA needs more specific rules and provisions to prevent over-removal of legal content. For instance, if the DSA provides incentives for content removal, it can lead to the over-removal of content as a way for social media platforms to shield themselves from liability.[10]

Secondly, the laws diverge significantly regarding the responsibilities placed on social media platforms and tech companies as intermediaries between information producers and citizens. Each country holds differing views regarding the role and obligations of tech firms in digital democracy. For instance, Germany's NetzDG and France's controversial "Avia Law," which is aimed at fighting hate speech, both mandate platforms to expeditiously take down

[1] Principal Researcher, Safer Internet Lab; Researcher, Department of Politics and Social Change, CSIS
[2] Deputy Executive Directors for Operations, CSIS
[3] Project Manager for Data Policy and Governance at Tifa Foundation; Research, Associate, Safer Internet Lab
[4] Independent Researcher; Research Associate, Safer Internet Lab
[5] Research Assistant, Safer Internet Lab
[6] Senior Researcher for the Citizen Lab, Munk School of Global Affairs & Public Policy, University of Toronto

[7] Aimée Vega Montiel and Emma Lygnerud Boberg. "Briefing note: Regulation, self-regulation and co-regulation in media and gender equality." International Media Support. 2021.

[8] Annegret Bendiek. "The Impact of the Digital Service Act (DSA) and Digital Markets Act (DMA) on European Integration Policy." Working Paper. 2021. https://www.swp-berlin.org/publications/products/arbeitspapiere/WP0121_Bendiek_Digital_Service_Act_and_Digital_Markets_Act.pdf

[9] Article 19. "At a glance: Does the EU Digital Services Act protect freedom of expression." February 11, 2021. https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/

[10] See, https://www.article19.org/wp-content/uploads/2022/02/A19-recommendations-for-the-DSA-Trilogue.pdf

manifestly illegal content when notified, imposing a heavy burden on companies to quickly moderate content or face fines.

In contrast, Brazil's Draft Bill No. 2630/20 does not mandate illegal content removal by platforms. Instead, it requires detailed transparency reports from social media companies about content moderation decisions and their systems.[11] Meanwhile, India's recently enacted IT Rules 2021 concerning digital media ethics have raised concerns about potentially allowing automated removal of content without human review, which could lead to violations of privacy and free speech.[12]

Thirdly, the laws diverge concerning the extent to which the general public, civil society groups and other stakeholders were involved in their development. The European Union's (EU) DSA and Brazil's Draft Bill No. 2630/2020 are often highlighted as positive examples of adequate consultancy and incorporation of inputs from non-governmental parties like academics, tech firms and civil society organisations during the drafting process.[13]

In contrast, Germany's NetzDG law and Singapore's Protection from Online Falsehoods and Manipulation Act (POFMA), which is an act to counter the proliferation of online falsehoods,[14] faced sharp criticism for the lack of public participation and input during the formulation of the regulations. It is necessary to involve the public, civil society groups, and technology companies in making and continually revising content moderation regulations so that there is public support for the rules and to address potential friction and resistance during the preemptive implementation stage.

Fourthly, the regulations analysed in this report are primarily national or regional laws in the case of the EU. However, the internet and social media networks operate globally with data that frequently crosses national borders. Thus, the enforceability of these laws is limited only to their country or region of origin. The EU's DSA represents an ambitious attempt to expand content moderation regulations across the region to achieve more harmonisation, but the EU still struggles to effectively govern platforms operating across multiple countries because of the subsidiarity to domestic law that might cause differences in the enforcement practices.[15]

Recognising the global nature of the internet, lawmakers need to pragmatically consider exactly how social media regulations can be enforced over multinational tech giants. Moreover, large, well-established multinational platforms tend to possess far greater resources, existing compliance processes, and public visibility than small regional startups and platforms. Thus, the compliance levels with regulations may vary widely across different sizes and types of platforms.

Fifthly, provisions for transparency and appeals in content moderation decisions differ markedly across the laws in focus. For instance, Brazil's Draft Bill 2630/20 contains detailed measures aimed at ensuring due process in content takedowns and account suspensions.[16] Similarly, a major focus of the EU's DSA is mandating various transparency reporting

[11] Draft Bill No. 2630 of 2020. Accessed from: https://cyberbrics.info/wp-content/uploads/2021/06/Brazilian-Fake-News-Draft-Bill-no.-2.630-of-2020.pdf

[12] Association for Progressive Communications (APC). "Proposed amendments to IT Rules in India threaten freedom of expression and privacy beyond borders." 2022. https://www.apc.org/en/press/proposed-amendments-it-rules-india-threaten-freedom-expression-and-privacy-beyond-borders

[13] Asha Allen. "CDT Europe Responds to European Commission Public Consultation on Templates for Transparency Reports Under EU Digital Services Act." 2024 and Talez Tomaz. "Brazilian Fake News Bill: Strong Content Moderation Accountability but LImited Hold on Platform Market Power." Journal of the European Institute for Communication and Culture, Vol. 30, No. (20). 2023.

[14] See, https://www.pofmaoffice.gov.sg/#:~:text=POFMA%20helps%20protect%20the%20Singapore,link%20to%20the%20Government's%20clarification.

[15] Cairán Seán Hennessy. "Addressing Disinformation through the DSA and CPD: Protecting democracy in the EU digital space." 2023.

[16] See, https://cyberbrics.info/wp-content/uploads/2021/06/Brazilian-Fake-News-Draft-Bill-no.-2.630-of-2020.pdf

requirements for platforms.[17] Meanwhile, Singapore's POFMA notably lacks transparency provisions and empowers government ministers to order social media platforms and sites to remove content or posts deemed false or misleading by the Singaporean government,[18] without requiring transparency or oversight safeguards.[19]

Transparency and appeals mechanisms for regular citizens and social media users deserve significant attention in formulating regulations focused on moderating online disinformation, especially around elections and political campaigns. The majority of social media users are also voters and politically engaged citizens who must have recourse to easily appeal if their content is removed. Platform accountability can also be enhanced by requiring regular detailed transparency reports from social media companies regarding content removals and account suspensions.

In summary, while most countries agree on addressing online harms like disinformation, they differ significantly around developing optimal approaches for regulating global technology firms, balancing security imperatives with free speech protections, and guaranteeing due process rights for average platform users and online citizens. The EU's DSA represents one of the more comprehensive and measured legal models proposed thus far. Yet, the enduring challenge of enforcing regulations on cross-border internet companies and platforms remains a central regulatory dilemma worldwide.

The Self-Regulation Approach

Citing the flaws of state-driven regulation, some major tech companies uphold a self-regulation approach to reviewing content moderation decisions, such as Meta's Oversight Board, TikTok's Safety Advisory Council, and X's (formerly known as Twitter) Trust and Safety Council. However, in 2022, the Twitter Trust and Safety Council was shut down, resulting in the absence of external insights into the platform.

Meta Oversight Board was established in 2020 as an independent body mandated to ensure that people's right to freedom of expression online is protected on the platform.[20] In this sense, the board offers an equal opportunity for users to appeal Meta's content decisions. The Board will review and decide what content to take down or leave up according to the platform's stated values and policies.[21] This decision can serve as a precedent for future content moderation decisions in Meta. Aside from reviewing individual cases, the Board can also accept requests from Facebook and Instagram to issue recommendations on its Content Policies. While the Boards' decisions for individual cases are binding, their recommendations for Meta are not.[22]

The Board consists of 22 members from diverse backgrounds at the time of writing to bring in people with various expertise and perspectives that can reflect on regional and local context-specific moderation issues.[23]

TikTok's Safety Advisory Councils consisted of members from industry and NGOs. These

[17] European Commission. "Very Large ONline Platforms and Search Engines to publish first transparency reports under the DSA." PRess Release. October 26, 2023.

[18] Kai Xiang Teo. "Civil Society Responses to Singapore's Online "Fake News" Law." International Journal of Communication. 2021.

[19] Chen Siyuan and Chia Chen Wei. "Singapore's Latest Efforts at Regulating Online Hate Speech: A Perspective from International Law and International Practices." 2019.

[20] Oversight Board. "Ensuring respect for free expression, through independent judgment." Accessed from: https://www.oversightboard.com/

[21] Oversight Board. "Ensuring respect for free expression, through independent judgment."

[22] Meta Transparency Report. "Oversight Board recommendations." December 21, 2023. https://transparency.fb.com/en-gb/oversight/oversight-board-recommendations/

[23] Oversight Board. "Meet the Board." Accessed from: https://www.oversightboard.com/meet-the-board/

councils work to develop forward-looking policies, product features, and safety processes that are informed by a diversity of perspectives, expertise, and lived experiences brought into the platform by council members.[24] To date, TikTok has established six Regional Safety Advisory Councils in Asia Pacific, Brazil, Europe, Latin America, MENAT (Middle East, North Africa, Turkey), and the US, known as the Content Advisory Council, and the company aims to expand their regional presence. Focusing on regions allows the council to create solutions based on a more targeted and responsive approach to safety while enabling the platform to keep up with the latest developments in each region. Through this regional approach, experts in the Councils collaborate to address problems in safety-related topics such as youth safety, free expression, and hate speech.[25]

X's now-dissolved Trust and Safety Council was created in 2016 as an advisory group consisting of around 100 volunteers such as independent civic leaders, activists, and academics.[26] The group provided expertise and guidance to combat a wide range of harmful content and safety issues, such as hate speech, harassment, and child exploitation, to name a few. In 2022, Twitter's owner, Elon Musk, disbanded the Trust and Safety Council, believing it was no longer the best structure to bring external insights to the platform's product and policy development. Although Musk had announced the plan of creating another content moderation council, this time under the leadership of X CEO Linda Yaccarino, it may be the case that the process will incorporate less input from outside experts in the future.[27] The absence of checks and balances in content moderation in X raises concerns regarding the safety and well-being of the users.

The presence of a platform-associated, independent/quasi-independent supervisory body may have its merits, such as:

1. **Increased transparency**. Boards can provide more visibility into content moderation policies, decisions, and disputes than platforms. Furthermore, Boards can highlight ambiguities and inadequacies of platforms' Community Standards.[28] Boards can also formalise the process of reviewing platforms' Community Standards or considerations that were not made publicly available when moderation decisions were made.
2. **Expertise.** Boards can leverage outside expertise from various fields to evaluate complex content issues. This expertise can complement internal platform knowledge in addressing safety-related problems.
3. **Advisory or guidance**. Boards can make non-binding policy recommendations to platforms to improve content rules and processes. The Oversight Board's decisions, investigations, and findings, for example, can provide Meta with insights to address blindspots in their content moderation decisions.
4. **Precedent.** Boards' past decisions on certain cases can serve as precedent for future content moderation for identical or similar content.
5. **Channel users' voice**. Boards allow users to appeal to content moderation decisions that affect them. Additionally, regional representation on these boards could improve users' access and awareness to appeal moderation decisions.
6. **Experimentation.** Self-regulation in the form of oversight boards allows platforms flexibility to try different models, adapt

[24] "Engaging Our Advisory Councils." TikTok, August 29, 2023. https://www.tiktok.com/transparency/en/advisory-councils/.

[25] TikTok. "Engaging Our Advisory Council." Accessed from: https://www.tiktok.com/transparency/en/advisory-councils/

[26] Amnesty International. "Global: Twitter's decision to disband safety council threatens wellbeing of users." December 13, 2022. https://www.amnesty.org/en/latest/news/2022/12/global-twitters-decision-to-disband-safety-council-threatens-wellbeing-of-users/

[27] "Twitter Dissolves Trust and Safety Council." The Washington Post, December 12, 2022. https://www.washingtonpost.com/technology/2022/12/12/musk-twitter-harass-yoel-roth/.

[28] David Wong and Luciano Floridi. "Meta's Oversight Board: A Review and Critical Assessment." *Minds and Machines* 33 (October 24, 2022): 261–84. https://doi.org/10.1007/s11023-022-09613-x.

them, and innovate before regulation is imposed.

7. **Global scale**. Platforms can more easily create oversight boards that span multiple jurisdictions than governmental bodies can. Furthermore, the translation of the platform's Community Standards to various home languages in Asian countries can be promoted to reach a wider audience, encourage compliance with their policies, and support their right to free expression.
8. **Speed.** Boards can respond more rapidly and be more nimble to emerging content issues than slower-moving governmental processes.
9. **Costs.** Self-regulation avoids the costs of developing new public regulatory bodies and processes by states to oversee platforms.

Nonetheless, the merits are coupled with its limitations, namely:

1. **Limited jurisdiction and authority** - Since oversight boards can only rule content moderation decisions on specific individual cases or posts. Aside from that, although the boards can issue recommendations for the platform's content policies, the platform can choose to disobey and not respond to them.[29] Therefore, the boards often have a narrow scope that circumscribes its impact. The Boards cannot also mandate changes to broader platforms' internal rules and features, such as algorithms, advertising systems, or data collection, resulting in the restriction of the Board to confront more systemic issues.
2. **Non-binding policy recommendations** - Policy suggestions from oversight boards are not binding for platforms. Therefore, the Boards cannot offer enforceable practical guidance, as they cannot compel platforms to change their underlying content moderation systems and policies. Therefore, as mentioned earlier, the platform cannot respond to the board's policy recommendations.[30]
3. **Limited reach and scale.** Oversight boards can only review a small fraction of platforms' content decisions. Reflecting the "quality over quantity" approach, it is unlikely that every user who has submitted an appeal will have their case reviewed by the board.
4. **There is a lack of transparency on boards**.

   There needs to be more visibility into how the boards operate, how cases are selected, and how competing considerations are weighed.
5. **Questionable independence**.

   While intended to be independent, oversight boards are typically funded and created by the platforms themselves. For instance, Meta's commitment to provide ongoing financial support for the Meta Oversight Board# raises concerns about their true autonomy and ability to counter the platforms' oversight boards.
6. **Limited diversity**.

   In some platforms, the board still lacks diversity in membership and is especially underrepresented in Global South countries despite the region encompassing most of the platform's audiences. For instance, the Meta Oversight Board only has one member from the Southeast Asia region. The membership also lacks diversity in non-geographical aspects, such as LGBTQ or disabled communities, that may not account for the geographic conceptions of diversity.
7. **Unclear precedential value**.

   The degree to which board decisions set precedents for future cases on platforms is often ambiguous and understudied, as the platforms' determination is controlled or determined.

[29] David Wong and Luciano Floridi. 2022.

[30] Meta Transparency Center. "Oversight Board: Further asked questions." January 19, 2022. https://transparency.fb.com/en-gb/oversight/further-asked-questions/

The Co-Regulation Approach: Codes of Practice

As governments grapple with regulating content moderation on digital platforms, an alternative model is emerging in the form of voluntary Codes of Practice. These industry codes represent a collaborative approach between platforms and public agencies. Rather than top-down state-imposed laws, Codes of Practice enable flexible self-regulation with government oversight and input. The experiences of Australia and the EU highlight constructive lessons in calibrating this co-regulatory balance.

The Australian Code of Practice on Disinformation and Misinformation (ACPDM), developed by the industry association Digital Industry Group Inc. (DIGI), demonstrates both the potential and limitations of voluntary self-regulation.[31] The code emerged from a government inquiry recommending platforms to address disinformation concerns. However, the ACPDM itself was formulated by industry actors based on multistakeholder consultations and this collaborative effort secured industry buy-in. The code promotes transparency, empowers users through media literacy, disrupts economic incentives for disinformation, and enables research access. Platforms can tailor commitments to their services through an opt-in model.[32]

The flexible nature of the ACPDM facilitated rapid progress and responsiveness. After an initial six-week trial period, the code underwent revisions to incorporate stakeholder feedback. However, relying mainly on self-regulation meant that enforcement mechanisms needed to be improved. While the Australian Communications and Media Authority (ACMA), as the industry regulator, provides oversight, non-compliance has few concrete consequences as opposed to binding regulations such as the Australian Code of Practice on Misinformation and Disinformation, which sanctioned social media companies for a maximum of 10,000 penalty units ($2.75 million in 2023) or 2 percent of global turnover for corporations.[33] Other critiques included high harm thresholds and content exemptions that could allow significant misinformation to persist. Overall, the ACPDM demonstrated the capacity of collaborative Codes of Practice to drive voluntary progress but underscored enforcement risks with industry self-regulation.

Learning from these limitations, the 2022 EU Strengthened Code of Practice on Disinformation (EU Code) incorporated more accountability levers. It expanded the scope to include misinformation, information operations, and foreign interference alongside disinformation.[34] Detailed commitments and quantitative reporting requirements enhanced transparency in the Code's implementation. Crucially, the EU Code is tied to enforcement powers and penalties for non-compliance by large platforms (more than 45 million users in the EU) in the Digital Services Act. This oversight mechanism lent more weight to the voluntary commitments.

The EU Code also focused on demonetising disinformation through advertising restrictions, thereby disrupting the toxic economy of harmful content. Furthermore, the Code allowed vetted researchers access to data and required transparency around recommender systems and political advertising. Additionally, the Code facilitated cross-platform collaboration and information sharing, which is aimed at addressing cross-cutting dynamics. However, allowing signatories to the Code discretion over which commitments to adopt meant that there needs to be more consistent application. Reliance on self-reporting and limited enforcement capacity also persist as

[31] DIGI Australia. Australian Code of Practice on Disinformation and Misinformation, 2022.

[32] DIGI Australia. "Australian Code of Practice on Disinformation and Misinformation | Annual Report." 2023.

[33] See, https://digi.org.au/wp-content/uploads/2023/10/Final-submission-on-exposure-draft-of-Communications-Legislation-Amendment-Combatting-Misinformation-and-Disinformation-Bill-2023-1.pdf

[34] European Commission. "2022 Strengthened Code of Practice on Disinformation," 2022. https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

concerns. Overall, though, linking the voluntary code to legal enforcement powers was an innovative development.

The experiences of the ACPDM and EU Code highlight the constructive role that Codes of Practice can play in content moderation governance while understanding and navigating their inherent limitations. The collaborative approach secures industry investment in the processes and outcomes. This approach also allows policies to adapt as digital risks evolve faster than legal reform cycles. Furthermore, tailored platform opt-in models provide flexibility to address distinct platform needs and business models. At the same time, periodic multistakeholder reviews enable the updating of codes to incorporate ongoing learnings and stakeholder feedback. These strengths make the co-regulatory approach promising.

Nonetheless, Codes of Practice often lack concrete enforcement mechanisms over their signatories. Therefore, reliance on self-regulation alone risks causing social harm, especially if companies deprioritise issues that have no significant impact on their profits, public image, or reputation. Allowing some discretion over commitments can also produce inconsistent regulation, while the capacity for monitoring compliance and auditing self-reported data remains a systemic constraint. As such, while voluntary Codes enable valuable collaboration, they likely work best when paired with enforceable legal mandates.

Using the approach of hybrid co-regulatory models, which combine self-regulatory codes with statutory requirements, could provide a balance for oversight bodies. For instance, the UK model designates the Office of Communications (Ofcom) as the regulator responsible for monitoring and enforcing compliance with the Code of Practice on Disinformation.[35] Coupling voluntary codes with renewed mandates for digital regulators to oversee its implementation and use deterrence mechanisms for violations could strengthen platforms' accountability, including (1) understanding the harms, (2) assessing the risk of harm, (3) deciding measures, implementing and record; and (4) report, review and update.[36] This blended co-regulation approach reconciles the flexibility of collaborative industry codes with the need for public oversight and enforcement.

As governments explore regulatory models for fast-moving issues like content risks, Codes of Practice demonstrate promising collaborative governance. However, their voluntary nature necessitates pairing with enforceable state-based oversight to ensure accountability. Well-designed co-regulatory frameworks that balance industry self-regulation with statutory enforcement powers could become a policy innovation to manage emerging digital and social risks. Achieving this hybrid balance remains an ongoing challenge but could be the orientation for policy experiments with voluntary codes. As platform governance involves shared public-private responsibilities, establishing Codes of Practice that thoughtfully combine self-regulation with public accountability can enable this partnership and strengthen its impact.

Future Directions in Social Media Platform Governance

Of the three different governance models being examined, key differences include the actors involved (platforms only, governments, or collaborative efforts), the mechanisms used (content policies, laws, or codes of conduct), and the scope (national laws vs. regional or international agreements). These approaches also differ in terms of the obligations and accountability imposed on platforms, with co-regulation often seen as a middle ground

[35] Ofcom. "Insight for online regulation: A case study monitoring political advertising." 2021. https://www.ofcom.org.uk/__data/assets/pdf_file/0020/212861/tools-for-online-regulation.pdf

[36] Ofcom. "Quick guide to online safety codes of practice." 2023. https://www.ofcom.org.uk/online-safety/information-for-industry/guide-for-services/codes-of-practice

between self-regulation and external laws.[37] However, these approaches all face challenges in implementation and their efficacy.

The three models need not be competing but rather complementary in the following ways:

1. State regulations can provide a legal framework and enforcement mechanisms, while self-regulation and co-regulation allow for flexibility, adaptability, and industry collaboration.
2. Co-regulation through Codes of Practice can bridge the gap between state regulation and self-regulation by combining the benefits of both approaches.
3. Hybrid co-regulatory models that pair voluntary codes with enforceable state-based oversight can ensure accountability and strengthen the impact of content governance initiatives.
4. Collaboration between governments, platforms, and other stakeholders can lead to more comprehensive and effective content governance strategies that address the complex challenges posed by the internet's global nature and evolving digital landscape.

In conclusion, a complementary approach that leverages the strengths of each model while mitigating their limitations could be the most effective way forward in governing social media content and addressing the challenges of online misinformation and disinformation.

[37] Michael Latzer, Natascha Just, and Florian Saurwein. "Self- and co-regulation: Evidence, legitimacy, and governance choice." In Book: Routledge Handbook of Media Law (pp.373-397), 2013.

Safer Internet Lab

saferinternetlab.org

Jl. Tanah Abang III no 23-27
Gambir, Jakarta Pusat. 10160

Find Us On

CSIS Indonesia | Safer Internet Lab