

LLMs are the Third Wave of Automated Content Moderation

Alex Engler¹ and Prithvi Iyer¹

¹ The University of Pennsylvania, Penn Center on Media, Technology, and Democracy

Abstract. Over the past two decades, online platforms have increasingly deployed algorithmic systems to make content moderation decisions, first through hashing algorithms and later using machine learning methods. Today, online platforms are increasingly exploring using large language models (LLMs), marking a third wave of automated content moderation. LLMs can be given prompts in plain language, are flexible to new policies and languages, and have some performance advantages. Adoption is being accelerated by falling costs, deep cuts to trust and safety teams, and public policy: takedown mandates measured in hours rather than days function as de facto automation requirements. Yet LLM moderation performs worse in low-resource languages, carries the political constraints of the models' origins, and reaches inconsistent decisions on identical content. We offer nine recommendations for platforms, regulators, and researchers.

Keywords: Content Moderation. Artificial Intelligence. Trust & Safety. Platform Governance. Large Language Models.

1 Introduction

Over the past two decades, online platforms have increasingly deployed algorithmic systems to make content moderation decisions, first through hashing algorithms and later using machine learning methods. Many large platforms automate over 90 percent of moderation actions. Today, online platforms are increasingly exploring using large language models (LLMs), marking a third wave of automated content moderation.

LLMs are not yet wholly replacing established algorithmic approaches, but they are tackling novel and challenging moderation decisions, which were previously often handled by individuals at trust and safety teams. This change is primarily being driven by technological advances of LLMs, the economic incentives of the online platforms, and by global digital policymaking. Content moderation is notoriously complex and nuanced: moderate too little and users face unchecked hate speech, misinformation, and other harmful content; moderate too much and platforms undermine freedom of expression and access to information. The scientific evidence and technical characteristics of LLMs demonstrate a mix of promise and peril from this third wave of automation—one that will shape the next generation of online platform governance. We conclude with nine essential recommendations in response, including calling for

expanded transparency and explanations; robust testing and new technical best practices; easier appeals and stronger autocratic safeguards; as well as support for the open-source community and an investment in multilingualism.

2 An Abbreviated History of Automated Content Moderation

In the early 2010s, online platforms primarily relied on human content moderators to manually review large volumes of user-generated content. Often performed by people in the Global Majority, this work comes with significant psychological toll and poor working conditions. These conditions persist, with recent research highlighting how moderators of non-English languages are routinely overworked, underpaid, and assigned content [1] in dialects they don't natively speak. However, the paradigm of content moderation has significantly moved toward automation, especially in the last ten years.

Moving past simple keyword filters, companies were experimenting with algorithmic content moderation at least as far back as 2007 [2]. For example, in the last three months of 2025, 98 percent of videos and 99.8 percent of comments removed [3] from YouTube were taken down by automated content moderation systems. In the latter half of 2025, TikTok's European Union operation used an automated system with no human review for 94 percent of content removals [4]. Until the past few years, online platforms have primarily relied on two computational techniques, paired with human review: matching algorithms and predictive models.

Matching algorithms compare newly posted content to see if it is identical or highly similar to content that has already been flagged in an existing database. One prominent form of matching is hashing, a type of "fingerprinting" in which an algorithm converts a piece of content into a unique numerical value, allowing platforms to detect identical or visually similar content at scale. A key limitation of matching algorithms is that they can only assess content that has already been identified as harmful by a human reviewer, not previously unseen content. Advances in hashing algorithms, especially perceptual hashing [5], do enable the identification of content that has been minorly altered.

Predictive models, by contrast, use machine learning methods to estimate the likelihood that a piece of content constitutes a policy violation. Human reviewers first label examples of content as consistent with, or in violation of, a platform's content policy; data scientists then train a model to estimate whether new content violates it. Because policies cover many topics with different criteria, platforms deploy multiple predictive models [6], one for each category, such as "hateful conduct" or "promotion of terrorism." The labels used to train these models are the product of subjective human judgment. A predictive model is therefore best understood not as producing an objectively correct answer, but as roughly reflecting the opinion [6] of the labelers that the model is emulating.

Predictive models have been fairly effective in some circumstances, such as classifying toxic comments with about 90 percent accuracy [7], albeit this is working in the easiest language (English) on an extensively studied problem. In 2017, Google Jigsaw and Wikimedia released the influential Perspective API, which used predictive models

to analyze posts for toxicity. While the API showed high accuracy for toxicity classification in testing, once released, users reported [2] numerous examples of “over- and under-zealous toxicity predictions.” One analysis found that four times as many German-language posts [8] would be moderated, relative to English-language posts.

While they are more flexible than matching algorithms, predictive models are also limited in their ability to classify novel and evolving content. Still, both techniques remain effective at vast scale and are essential to systems identifying terrorist content and child sexual abuse material (CSAM). But platforms must also address novel and fast-changing content where the language around what is harmful is constantly evolving: enter large language models.

3 The Emerging Role of LLMs in Content Moderation

Over the past several years, internal teams at online platforms and independent companies have begun to explore and implement LLMs for content moderation. Systematic reviews have demonstrated that LLMs outperform the prior generations of content moderation algorithms in some domains, for instance in hate speech detection [9].

3.1 Flexible, adaptable, and multi-lingual

The original applications of LLM-based content moderation, such as Meta's LlamaGuard and Google's ShieldGemma, used fixed definitions of content violations that the model was trained or adapted to identify. By contrast, more recent LLM-based systems enable users to write their own policies for allowable content. This new family of bring-your-own-policy (BYOP) tools includes OpenAI's gpt-oss-safeguard [10] or Zenropi's CoPE-B [11], which allow online platforms to write content policies in plain language and give them to the model. This also enables the sharing of effective LLM policies, for instance, those written by the Christchurch Call on violent extremism. Plain-language policy editing also reduces, though does not eliminate, the engineering work of adjusting a deployed policy—an important consideration for under-resourced trust and safety teams. The BYOP approach is therefore especially beneficial for quickly evolving discussions and virally spreading content.

Multi-lingual fluency is another significant advantage for LLM-based moderation, with some LLMs supporting 80 or more languages. In one study [12], a fine-tuned Llama model matched or outperformed prior methods for detecting predatory and abusive conversations in English, Roman Urdu, and Urdu. Unsurprisingly, this benefit is neither automatic nor guaranteed. An analysis of LLM content moderation for cross-cultural English language hate speech found a clear bias towards Anglosphere English [13] relative to English spoken in South Africa and Singapore. A systematic review [9] of LLM hate speech detection found “while LLMs have demonstrated remarkable success in resource-rich languages, their cross-lingual and multilingual performance often falls short.”

3.2 Cost considerations amidst widespread layoffs

Deploying a high-quality LLM classifier for every piece of content, especially on larger platforms, would be extremely expensive [11] and prohibitively slow. This is especially true for reasoning models [14], which use much more computing. However, there are paths to more affordable uses of LLMs. Recent work [15] demonstrates that smaller BYOP models can also classify content according to policy instructions delivered as prompts, without retraining each time a policy changes.

Hiring and training human moderators worldwide is also expensive. These teams have seen dramatic cuts, exacerbated by political pushback against online safety work. After Elon Musk took over Twitter in 2022, the company laid off one third [16] of its trust and safety staff. In August 2025, TikTok also announced layoffs of hundreds of content moderators as it expanded its use of AI for content moderation [17].

In March 2026, Meta announced a years-long rollout of AI systems [18] to handle content enforcement tasks, reducing its reliance on third-party vendors and contractors. The largest platforms are already investing heavily in LLMs, so automating moderation is an obvious way to help justify that expenditure.

3.3 Handling novel and complex content

Established algorithms struggle with novel or subtle forms of banned content. It is for these challenging edge cases that LLMs seem promising [19] and worth their cost. LLMs, especially those with reasoning capabilities, are far better at contextual understanding of natural language [20] relative to prior algorithms, including the ability to interpret sarcasm, implicit meaning, and culturally embedded references.

LLMs may prove especially valuable for escalating (that is, notifying human moderators for further deliberation) challenging cases that require contextual understanding. One proposed hybrid system [19] has LLMs decide clear violations while flagging edge cases and emerging threats to human moderators, whose decisions then feed back into training data and policy refinement.

Additionally, when prompted, LLMs may be able to do more than escalate and classify content [20]: they can offer explanations for their reasoning by pointing to specific rules or sections of relevant content moderation policies. A stronger body of research is necessary to draw strong conclusions about these LLM-based explanations. However, when a user is given an explanation as to why their content was taken down, there is some evidence this will decrease future violations [21].

4 Repercussions of LLM Content Moderation

Despite their advantages over prior algorithmic approaches, LLMs in content moderation have also demonstrated shortcomings regarding fairness, bias, and free expression. While LLMs can be adapted more effectively to other languages and dialects than prior algorithms, they still perform worse [22] in low-resource languages, that is, in languages with minimal or low-quality training data. Most current language models reflect

the world as seen through Anglo-centric and predominantly North American texts, introducing biases against languages and cultural perspectives not represented in training data. This creates an equity problem for content moderation: communities already underserved are more likely to have their content incorrectly flagged, or harmful content directed at them go undetected. Data-donation programs and community-led initiatives that build models by and for the Global Majority offer one route to closing this gap.

These efforts may also help address a concern raised by the European Center for Not-for-Profit Law, which argues that LLMs deployed into online platforms may promote an “algorithmic monoculture [23]” wherein a small number of foundation models can dictate or strongly influence global online speech norms. This concern has been borne out in evidence—a comprehensive audit [24] found that popular LLM moderation tools (including from Google, Anthropic, and OpenAI) suppress identity-related text two or three times more frequently than text that is not identity-related.

LLM chatbots also suppress or censor some political topics, which would likely carry over to moderation decisions. Prompted with 145 Chinese political topics [25], models developed in China refused around one third of prompts, versus zero to three percent for non-Chinese models. In some cases this may reflect intentional censorship, but in others, especially authoritarian countries like China and Russia, longstanding state control over the local media ecosystem has biased [26] the data the LLMs are trained on. Some AI companies, including Anthropic, have acknowledged this issue.

Beyond explicit censorship, there is also the risk of “algorithmic arbitrariness [27],” with one study finding that multiple models (each performing equally well on average) can reach conflicting decisions on the same piece of content, depending on even trivial aspects of model design. Recent work [28] provides further evidence for this concern. In the first large-scale comparative analysis of seven leading moderation systems, they find that identical content receives markedly different hate speech classifications depending on the LLM used. A related concern is that even sophisticated LLMs [29] struggle to distinguish between the use of harmful language and references to that harmful language. This is critical because refuting problematic claims, or counterspeech, often involves mentioning those problematic claims.

These concerns are not entirely novel: the 2025 joint declaration [30] by the United Nations, OSCE, OAS, and African Commission special rapporteurs on freedom of expression cautioned that the use of algorithms for content moderation can lead to over-removal, discrimination, and censorship. As LLMs take on a broader role in content moderation, it is not yet clear whether they will be a tool to address these algorithmic shortcomings, or will ultimately exacerbate them.

5 Public Policy is Encouraging LLM-Automated Content Moderation

Platforms are likely to adopt LLMs for their flexibility and efficacy on novel content, especially as cost declines [31]—but they are also responding to public policy. In legislating content takedown requirements with short time windows, governments are creating de facto automation requirements, encouraging the adoption of LLMs. Meta’s

transparency report notes that, while uncommon, “a country’s law may obligate us to automatically restrict access to content, at scale in specific countries, based on local law requirements” [32].

One of the first major laws [33] regulating online content moderation came in Germany’s 2017 NetzDG law, which compelled companies to remove “manifestly unlawful” content within 24 hours of a user complaint. The UK has similar laws: non-consensual intimate imagery will have to be taken down within 48 hours of notification [34]; while the Online Safety Act mandates [35] that platforms use systems and processes to “minimise the length of time for which any priority illegal content is present.” Since 2022, the European Union has required removal of terrorist content [36] within one hour of a government removal order.

In February, India’s Information Technology ministry passed amendments [37] to the IT rules that mandate social media platforms comply with content takedown notices within 3 hours. Meta’s compliance with such notices has risen sharply: it restricted 1,808 items in India under local law in 2022, 25,600 in 2024, and 69,400 in 2025 [32]. Amid ongoing repression of student protests, Indian agencies issued an average of 1,275 blocking orders a day between March and July 2026 [38], and Meta is resorting to broad automated censorship [39] to adhere to the Modi government’s strict content takedown mandates, likely motivated by the legal risk to the platform’s safe harbor protection. Vietnam’s latest amendments to its cyber law [40] also follow this trend. Platforms now have only 24 hours [41] to remove content that violates its broad mandate, and merely 6 hours for “urgent” cases. Meta, YouTube, and TikTok have complied in over 90% of cases [41] with the Vietnamese government’s takedown requests.

According to Freedom House, in at least 21 countries platforms were required, explicitly or through tight removal deadlines, to use automated systems [42] for content moderation. Governments are demanding faster, more comprehensive compliance, and we should expect continued automation as a result. As Apar Gupta, Executive Director of India’s Internet Freedom Foundation stated [43]: “Platforms, facing legal risk, will deploy error-prone automated systems, label widely, remove quickly, and prefer false positives to liability.”

6 Recommendations

We highlight nine recommendations made more essential by this third wave of automated content moderation.

Adapted Transparency Reporting—As part of transparency reporting, online platforms should provide, and technology regulators should request, information about the automated systems used for legally required content moderation, including whether they are matching algorithms, predictive models, LLM-based systems, or a combination thereof. For LLM-based moderation, it is also worth collecting which specific LLM is being used, how it has been adapted and deployed, and for what purposes, as performance varies extensively.

New Technical Best Practices—LLM developers should publish best practices for LLM use in content moderation, such as clear model specifications, system prompts

that balance content safety and free expression, and fine-tuning assisted by subject matter experts with appropriate cultural knowledge. For high-stakes content, platforms should also explore ensembles of LLMs that vote on decisions to reduce arbitrary or erroneous decisions.

Localized Testing and Evaluation—For particularly critical content (e.g. CSAM, terrorist content, calls to violence, etc), online platforms should perform, and regulators should require testing and evaluation of automated content moderation systems in local languages and dialects, evaluated against language-fluent, in-country, culturally attuned experts. Regulators should also perform independent evaluations and fund third party research, including longitudinal monitoring of LLMs.

New Possibilities for Explanations—Online platforms and independent researchers should investigate the merits and limitations of LLM-based explanations for content moderation decisions. If meaningful, these explanations might also provide insights to regulators through regulatory mandates such as Article 17 of the EU’s Digital Services Act [44].

More Essential Appeals—As content moderation becomes progressively more automated, a path to a human appeal becomes ever more essential. Online platforms should maintain, and policymakers should require, appeal processes that are obvious, accessible, and lead expeditiously to a review by an appropriately trained human.

Encouraging Open-Source LLMs—The emergence of LLMs in content moderation is another reason to support open-source LLMs. A vibrant open-source ecosystem offers platforms more competition and optionality in building effective and equitable moderation systems.

Invest in improving LLMs for Low-Resource Languages—Online platforms, AI companies, researchers, and philanthropies should invest in the data infrastructure underpinning multilingual moderation, including data-donation programs and community-led model development.

Stronger Safeguards from Autocratic Manipulation—BYOP LLMs may enable easier autocratic manipulation, since mandating a specific policy prompt is easier than demanding a bespoke algorithmic system. Online platforms should strengthen standards and mechanisms for refusing unlawful or unethical content moderation demands from autocratic regimes.

Refrain from Filtering Mandates—Governments have at times considered requiring online platforms to proactively identify and automatically filter certain types of content when it is uploaded [45]. LLMs do not significantly change the balance of costs and benefits of such mandates, and governments should continue to refrain from them.

Acknowledgments.

We are grateful for the considered feedback of Danaé Metaxa, Daphne Keller, and Matt O’Shaughnessy.

Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Elswah, M., Bhatia, A., Thakur, D.: Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages. Center for Democracy & Technology, Washington, DC (2025). <https://cdt.org/insights/content-moderation-in-the-global-south-a-comparative-study-of-four-low-resource-languages/>
2. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data Soc.* 7(1) (2020). <https://doi.org/10.1177/2053951719897945>
3. Google: YouTube Community Guidelines enforcement. Google Transparency Report. <https://transparencyreport.google.com/youtube-policy/removals>
4. TikTok: Digital Services Act: Our Sixth Transparency Report on Content Moderation in Europe. TikTok Newsroom (2026). <https://newsroom.tiktok.com/digital-services-act-our-sixth-transparency-report-on-content-moderation-in-europe?lang=en-150>
5. Zhou, Y., Li, X., Xiong, C., Yao, H., Qin, C.: A survey of perceptual hashing for multimedia. *ACM Trans. Multimed. Comput. Commun. Appl.* 21(7), 1–28 (2025). <https://doi.org/10.1145/3727880>
6. Chowdhury, N.: Automated Content Moderation: A Primer. Program on Platform Regulation, Stanford Cyber Policy Center. https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/automated_content_moderation_a_primer.pdf
7. Rahman, M.A., Nayem, A., Amjad, M., Siddik, M.S.: How do machine learning algorithms effectively classify toxic comments? An empirical analysis. *Int. J. Intell. Syst. Appl.* 15(4), 1–14 (2023). <https://doi.org/10.5815/ijisa.2023.04.01>
8. Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., Giordano, S.: Toxic bias: Perspective API misreads German as more toxic. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, pp. 1346–1357. AAAI Press (2025)
9. Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., Seals, C.: Hate speech detection using large language models: a comprehensive review. *IEEE Access* 13, 20871–20892 (2025). <https://doi.org/10.1109/ACCESS.2025.3532397>
10. OpenAI: Introducing gpt-oss-safeguard (2025). <https://openai.com/index/introducing-gpt-oss-safeguard/>
11. ROOST: Welcoming Zentropi’s CoPE-B-A4B to the ROOST Model Community (2026). <https://roost.tools/blog/welcoming-zentropi-s-cope-b-a4b-to-the-roost-model-community/>
12. Nguyen, T.T., Wilson, C., Dalins, J.: Fine-tuning Llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv:2308.14683* (2023)
13. Lee, N., Jung, C., Myung, J., Jin, J., Camacho-Collados, J., Kim, J., Oh, A.: Exploring cross-cultural differences in English hate speech annotations: from dataset construction to analysis. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4205–4224. Association for Computational Linguistics, Mexico City (2024)
14. Mitchell, M.: Artificial intelligence learns to reason. *Science* 387(6740), eadw5211 (2025). <https://doi.org/10.1126/science.adw5211>
15. Chakrabarti, S., Willner, D., Klyman, K., Saade, T., Capstick, E., Nong, S.: CoPE: a small language model for steerable and scalable content labeling. *arXiv:2512.18027* (2025)
16. eSafety Commissioner: Report reveals the extent of deep cuts to safety staff and gaps in Twitter/X’s measures to tackle online hate. Australian Government. <https://www.esafety.gov.au/newsroom/media-releases/report-reveals-the-extent-of-deep-cuts-to-safety-staff-and-gaps-in-twitter/xs-measures-to-tackle-online-hate>

17. Wall Street Journal: TikTok to lay off hundreds of U.K. content moderators in AI push (2025). <https://www.wsj.com/tech/tiktok-to-lay-off-hundreds-of-u-k-content-moderators-in-ai-push-d040ae21>
18. CNBC: Meta to cut back on third-party vendors in favor of AI for content enforcement (2026). <https://www.cnbc.com/2026/03/19/meta-cut-back-third-party-vendors-favor-of-ai-for-content-enforcement.html>
19. Hunsberger, A.: LLM Content Moderation: Implementation Guide for Trust & Safety Teams. Musubi Labs (2026). <https://www.musubilabs.ai/post/the-top-challenges-of-using-llms-for-content-moderation-and-how-to-overcome-them>
20. Chen, C., Qu, W., Su, S.-F., Feng, Y., Li, T.: A comprehensive review of LLM-based content moderation: advancements, challenges, and future directions. *Knowl.-Based Syst.* 330, 114689 (2025). <https://doi.org/10.1016/j.knosys.2025.114689>
21. Prosocial Design Network: Removal Explanations. <https://www.prosocialdesign.org/library/removal-explanations>
22. Pava, J.N., Meinhardt, C., Badi Uz Zaman, H., Friedman, T., Truong, S.T., Zhang, D., Cryst, E., Marivate, V., Koyejo, S.: Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts. Stanford Institute for Human-Centered Artificial Intelligence (2025). <https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts>
23. Wisniak, M.: Algorithmic Gatekeepers: The Human Rights Impacts of LLM Content Moderation. European Center for Not-for-Profit Law (2025). https://ecnl.org/sites/default/files/2025-04/ECNL_LLM_CM_Executive%20Summary_2025.pdf
24. Proebsting, G., Anigboro, O.I., Crawford, C.M., Metaxa, D., Friedler, S.A.: Identity-related speech suppression in generative AI content moderation. In: Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 185–217. ACM (2025). <https://doi.org/10.1145/3757887.3763010>
25. Pan, J., Xu, X.: Political censorship in large language models originating from China. *PNAS Nexus* 5(2), pgag013 (2026). <https://doi.org/10.1093/pnasnexus/pgag013>
26. Waight, H., Yang, E., Yuan, Y., Messing, S., Roberts, M.E., Stewart, B.M., Tucker, J.A.: State media control influences large language models. *Nature* 655(8123), 685–693 (2026). <https://doi.org/10.1038/s41586-026-10506-7>
27. Gomez, J.F., Machado, C.V., Paes, L.M., Calmon, F.P.: Algorithmic arbitrariness in content moderation. *arXiv:2402.16979* (2024)
28. Fasching, N., Lelkes, Y.: Model-dependent moderation: inconsistencies in hate speech detection across LLM-based systems. In: Findings of the Association for Computational Linguistics: ACL 2025, pp. 22271–22285. Association for Computational Linguistics, Vienna (2025)
29. Gligorić, K., Cheng, M., Zheng, L., Durmus, E., Jurafsky, D.: NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5942–5959. Association for Computational Linguistics, Mexico City (2024)
30. UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression, ACHPR Special Rapporteur on Freedom of Expression and Access to Information in Africa: Joint Declaration on AI, Freedom of Expression and Media Freedom (2025). <https://www.ohchr.org/sites/default/files/documents/issues/expression/statements/2025-10-24-joint-declaration-artificial-intelligence.pdf>

31. Xiao, C., Cai, J., Zhao, W., Lin, B., Zeng, G., Zhou, J., Zheng, Z., Han, X., Liu, Z., Sun, M.: Densing law of LLMs. *Nat. Mach. Intell.* 7(11), 1823–1833 (2025). <https://doi.org/10.1038/s42256-025-01137-0>
32. Meta: Content restrictions based on local law: India. Meta Transparency Center. <https://transparency.meta.com/reports/content-restrictions/country/IN/>
33. Gorwa, R.: Elections, institutions, and the regulatory politics of platform governance: the case of the German NetzDG. *Telecommun. Policy* 45(6), 102145 (2021). <https://doi.org/10.1016/j.telpol.2021.102145>
34. Department for Science, Innovation and Technology: Tech firms will have to take down abusive images within 48 hours under new law to protect women and girls. GOV.UK (2026). <https://www.gov.uk/government/news/tech-firms-will-have-to-take-down-abusive-images-within-48-hours-under-new-law-to-protect-women-and-girls>
35. Online Safety Act 2023, c. 50 (2023). <https://www.legislation.gov.uk/ukpga/2023/50>
36. European Parliament: New rules adopted for quick and smooth removal of terrorist content online. Press release (2021). <https://www.europarl.europa.eu/news/en/press-room/20210422IPR02621/new-rules-adopted-for-quick-and-smooth-removal-of-terrorist-content-online>
37. Ministry of Electronics and Information Technology: The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, as amended. Government of India (2026). <https://www.meity.gov.in/static/uploads/2026/02/550681ab908f8afb135b0ad42816a1c9.pdf>
38. The News Minute: 1 takedown order every 68 seconds: Union issued 1.95 lakh blocking orders in 5 months (2026). <https://www.thenewsminute.com/news/1-takedown-order-every-68-seconds-union-issued-195-lakh-blocking-orders-in-5-months>
39. The Hindu: Meta sets up automatic blocking for govt, police-referred content (2026). <https://www.thehindu.com/sci-tech/technology/meta-sets-up-automatic-blocking-for-govt-police-referred-content/article70901931.ece>
40. National Assembly of the Socialist Republic of Vietnam: Law on Cybersecurity No. 116/2025/QH15 (2025). <https://english.luatvietnam.vn/lawoncybersecurityno116-2025-qh15dateddecember102025ofthenationalassembly-422396-doc1.html>
41. Alam, K.: Vietnam tightens online speech controls as cyber law takes effect. *Tech Policy Press* (2026). <https://www.techpolicy.press/vietnam-tightens-online-speech-controls-as-cyber-law-takes-effect/>
42. Funk, A., Shahbaz, A., Vesteinsson, K.: Freedom on the Net 2023: The Repressive Power of Artificial Intelligence. Freedom House, Washington, DC (2023). <https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence>
43. Gupta, A., Pahwa, N.: The rise of infrastructure for digital censorship. Internet Freedom Foundation (2026). <https://internetfreedom.in/the-rise-of-infrastructure-for-digital-censorship/>
44. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act), Article 17. *Off. J. Eur. Union L* 277, 1–102 (2022). <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
45. Access Now et al.: Open letter to Members of the European Parliament on the proposed regulation on preventing the dissemination of terrorist content online (2019). <https://www.accessnow.org/wp-content/uploads/2019/02/Access-Now-Civil-Society-Letter-to-EP-Terrorism-Database.pdf>