TECHGLOBAL INSTITUTE

Drawing the Lines: Navigating the Content, Conduct, and Context of Borderline Content Moderation

Diyi Liu

Shahzeb Mahmood

Acknowledgement

Diyi Liu would like to acknowledge the support received from the Tech Global Institute in facilitating the research and development of this paper. The authors would like to acknowledge the interviewees who contributed their time and expertise to the research, as well as colleagues and others who provided feedback, assistance, and support throughout the project.

Introduction

In 2016, Facebook deleted a post by Norwegian writer Tom Egeland depicting the Pulitzer prize-winning photograph "The Terror of War" (also known as the "Napalm Girl"), for allegedly violating its rules on child nudity.[1] This moderation decision triggered global outrage.[2] In response, Facebook reversed its decision, noting its "status as an iconic image of historical importance" and introduced a newsworthiness allowance, which meant "allowing more items that people find newsworthy, significant, or important to the public interest — even if they might otherwise violate our standards."[3]

In an interview with CBS News, former YouTube CEO Susan Wojcicki walked through the determination of two borderline cases and its real-world consequences. Syrian war activists, for instance, criticized YouTube in 2017 for removing historical documentation of the conflict, arguing that the platform was "destroying digital evidence" crucial for atrocity trials.[4] During the interview,

Figure 1: Huỳnh Công Út captured the Pulitzer Prize-winning photograph of Phan Thị Kim Phúc fl eeing a napalm attack in 1972 for the Associated Press, widely considered as one of the defi ning photographs of the Vietnam War.

1 BBC News, "Facebook U-turn Over 'Napalm Girl' Photograph," *BBC News*, September 9, 2016, https://www.bbc.com/news/technology-37318040.

2 Levin et al., "Facebook Backs down from 'Napalm Girl' Censorship and Reinstates Photo," Technology, *The Guardian*, September 9, 2016, https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo.

3 Joel Kaplan and Justin Osofsky, "Input From Community and Partners On Our Community Standards," *Meta*, October 21, 2016, https:/about.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-st andards/. See also Meta, "Our Approach to Newsworthy Content," *Meta Transparency Center*, November 12, 2024, https://transparency.meta.com/features/approach-to-newsworthy-content/.

4 TRT Global, "TRT Global - Activists Accuse YouTube of Destroying Digital Evidence of Syria War," March 8, 2021, https://trt.global/world/article/12873925.

Wojcicki explained that while the content was undeniably graphic and involved violence, users who posted the content intended to expose atrocities, and the context served clear public interest in documenting war crimes. Ultimately, YouTube decided to keep the content accessible. The second case involved what appeared to be World War II historical footage. Yet the video was removed because of the presence of "1418," a numerical code used by white supremacists to identify one another. As Wojcicki noted in the interview, "there's quite a lot of context that goes into every single video to be able to understand what they are really trying to say with this video."[5]

Far from being isolated incidents, these episodes exposed some fundamental tensions at the heart of content moderation by social media platforms. Specifically, how do these platforms define and apply policy allowances and exemptions, particularly in balancing consistent enforcement with contextual judgment in assessing material that is historically, culturally, and journalistically significant? Should, for instance, a platform's nudity policies be applied to an image of historical significance? Broadening the scope of discussion, should a platform's hate speech, violence, and incitement policies be subjected to exemptions to allow civilians in conflict regions to document their experiences, which may subsequently prove vital for atrocity trials, or otherwise treat such content as political expression that holds a broader remit of protection than normal speech?[6]

Most of these questions lack straightforward answers, yet these are the challenging decisions social media platforms increasingly face: how to moderate at scale content that exists in a space of contextually contingent interpretive ambiguity with respect to existing speech rules, which has come to be known as "borderline content". Such cases fall into what legal scholars and practitioners increasingly describe as **"lawful but awful"** or **"hard-to-adjudicate"** content areas, where content might not clearly violate specific laws or platform content policies but still raises serious concerns about harm, ethics, or social impact.[7] Unlike unambiguous policy violations that can

5 Leslay Stahl, "Is YouTube Doing Enough to Fight Hate Speech and Conspiracy Theories?," CBS News, December 1, 2019, https://www.cbsnews.com/news/is-youtube-doing-enough-to-fight-hate-speech-and-conspiracy-theories-60-minutes-2019-12-01/.

6 Munsif Vengattil and Elizabeth Culliford, "Facebook Allows War Posts Urging Violence against Russian Invaders," Reuters, March 11, 2022, https://www.reuters.com/world/europe/exclusive-facebook-instagram-temporarily-allow-calls-violence-against-russians-2022-03-10/.

7 Daphne Keller, "Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users," *University of Chicago Law Review Online*, July 7, 2022, https://lawreview.uchicago.edu/online-archive/lawful-awful-control-over-legal-speech-platforms-governments-and-internet-users; Anastasia Kozyreva et al., "Resolving Content Moderation Dilemmas between Free Speech and Harmful Misinformation," *Proceedings of the National Academy of Sciences of the United States of America* 120(7) (2023): e2210666120, https://www.pnas.org/doi/10.1073/pnas.2210666120.

be swiftly actioned, borderline content often generates substantial disagreement among stakeholders about whether and how it should be governed, and by whom. These cases frequently occupy a contested regulatory and normative terrain, where competing interpretations, values, and assessments of potential harm complicate moderation decisions.

This research seeks to advance our understanding of borderline content moderation. Moving beyond clear-cut definitions, it examines the interplay of three factors that make certain cases difficult to adjudicate.

First, the **content** itself may possess visual, textual, or audio characteristics that appear to approach policy boundaries without clearly violating them. Even within ostensibly well-defined categories such as child sexual abuse material, policy experts and practitioners report encountering borderline scenarios that require interpretive judgment.

Second, the **conduct** surrounding the content — including actions, intentions, and patterns of behaviour associated with content creation, dissemination, or engagement — may generate ambiguity. Content that is permissible in isolation may become problematic when assessed alongside patterns of behavior, coordination, or signals of harmful, deceptive, or otherwise problematic intent or activity.

Third, the **context** in which content is created, shared, and interpreted shapes moderation decisions. Factors such as newsworthiness, artistic expression, and cultural, legal, or historical setting can alter how content is understood. Content deemed borderline in one jurisdiction may be acceptable in another, creating additional challenges for global platforms.

A meme circulated during a visit by Indian Prime Minister Narendra Modi to Bangladesh in March 2021, showing him alongside then Bangladeshi Prime Minister Sheikh Hasina Wazed walking backstage through curtains and featuring a caption asking, "What will happen behind the curtain?" accompanied by a wink emoji. Viewed solely as content, the meme consisted of an image of two public figures with a mocking caption, which, in isolation, could plausibly be interpreted as political humour, satire, or criticism and therefore might not clearly violate platform policies.

Viewed through the lens of conduct, however, additional considerations emerge. For some audiences, the caption carried a sexual innuendo suggesting an improper relationship between the two political leaders. Multiple accounts posted this meme and other content making similar sexually suggestive remarks about Hasina in relation to multiple male political figures over several years — and moderators reviewing this content might interpret it not as isolated satire but as part of a broader pattern of gendered and targeted harassment. The content itself remains unchanged, but user conduct and behavioural history alter how it is evaluated.

When refracted through the lens of context, the meme raises a different set of concerns. First, the insinuation relied upon allegations of sexual impropriety and sexual indiscretion directed at a female political leader, in a society where questions of sexual morality, reputation, and honour remain highly salient. The content therefore may carry consequences beyond ordinary political criticism and risk reinforcing gendered stereotypes among conservative religious and political constituencies. Second, and more significantly, the meme circulated within a highly charged geopolitical and domestic political context — relations between Bangladesh and India have long been marked by debates over sovereignty, border management, water-sharing disputes, economic dependence, and perceptions of Indian influence in Bangladeshi politics. Modi's visit occurred amid widespread protests by opposition groups. Violent clashes associated with the protests resulted in the deaths of around a dozen people and generated significant political tensions. Within this context, the meme's suggestion that Sheikh Hasina was "in bed" with Modi could be interpreted not merely as a personal insult but as an allegation of political subservience and obsequiousness. The sexual metaphor thus intersected with existing nationalist grievances and anti-government sentiment, increasing the possibility that the content may have contributed to online hostility and harassment, as well as offline tensions.

This example illustrates how borderline content may remain lawful and seemingly permissible when assessed solely on the basis of its textual and visual characteristics, yet become considerably more difficult to adjudicate once moderators take account of user conduct, political context, and the potential for harmful interpretations. The challenge for platforms lies precisely in assessing the content-conduct-context matrix and determining how much weight should be given to these surrounding factors when making moderation decisions.

Through policy analysis and case studies of major social media platforms and in-depth interviews with content moderators, policy specialists, and trust and safety practitioners, the research demonstrates that the contested nature of borderline content emerges from the interaction between content, conduct, and the social contexts. It traces how borderline content moves through escalation processes and reveals the complex decision-makings and institutional arrangements that govern these difficult determinations.

Perhaps most significantly, this work exposes **critical power imbalances** in current moderation practices. Decision-making authority typically resides with Western, English-speaking corporate senior executives, while those with local cultural and linguistic expertise, often based in Majority World countries, function primarily as reviewers and consultants without meaningful influence over policy debates or moderation decisions. These structural inequities are compounded by inadequate training resources, exploitative vendor relationships, and insufficient mechanisms for managing stakeholder disagreement.

The findings point toward several pathways for improvement: developing more contextually responsive governance frameworks that accommodate local interpretation, creating better coordination mechanisms among diverse stakeholders, establishing enhanced training protocols for interpretively complex cases, and, more importantly, restructuring decision-making processes to center rather than marginalize expertise from the Majority World. Ultimately, this work offers a foundation for more effective and context-sensitive content governance in a complex global landscape, one shaped by the moderation of not only user-generated but increasingly content generated by artificial intelligence (AI), where the boundaries of appropriate content continue to evolve, and the stakes of getting these decisions right have never been higher.

TECHGLOBAL INSTITUTE

Background

Content moderation on global social media platforms involves difficult trade-offs and moral dilemmas. It operates within overlapping layers of normative principles and policies that create both opportunities and constraints for platform decision-making.

Firstly, companies develop comprehensive content policies that determine not only what content is permissible, but increasingly what should be subject to intermediate interventions, reduction, de-amplification, labeling, or removal from recommendation algorithms or from the platform itself. The wide spectrum of moderation enforcement reflects platforms' growing recognition that not all potentially harmful or sensitive content requires the same level of intervention, and that moderation measures should be calibrated according to the necessity and proportionality of the action taken.

Despite their wide-ranging implications across jurisdictions, however, these policies are largely curated unilaterally and internally by platforms — primarily within the normative and legal affordances of the American First Amendment free speech tradition — without meaningful public consultation or transparent decision-making processes. This could lead to the apparent homogenization of rules across diverse legal, cultural, and social contexts, while raising questions about the extent to which moderation frameworks allow for contextually attuned assessments.

Secondly, global platforms must navigate diverse jurisdictional requirements while maintaining operational coherence. The regulatory landscape has intensified significantly in recent years, with initiatives like Australia's *Online Safety Act 2021*, the European Union's *Digital Services Act 2022*, United Kingdom's *Online Safety Act 2023*, or Sri Lanka's *Online Safety Act, 2024*, and various national laws creating complex compliance matrices for platforms operating across multiple jurisdictions.[8] It remains unclear, however, to what extent these national laws are accommodated within privately curated platform rules, and how, whether, and under what circumstances such legislations are enforced domestically or applied extraterritorially, and how platforms mediate the tensions inherent in navigating divergent normative, legal, and cultural approaches across jurisdictions.

8 Yasmin Afina et al., *Towards a Global Approach to Digital Platform Regulation: Preserving Openness amid the Push for Internet Sovereignty*, Digital Society Initiative (Chatham House, 2024), https://chathamhouse.soutron.net/Portal/Public/en-GB/RecordView/Index/203893.

Thirdly, international human rights standards serves as foundations for many rights-based approach, particularly the International Covenant on Civil and Political Rights (ICCPR) and the UN Guiding Principles on Business and Human Rights (UNGP), which have become central not only to various multistakeholder initiatives like the Global Network Initiative (GNI) and the Global Internet Forum to Counter Terrorism (GIFCT), but also received formal and public commitments by most major platforms to adhere to the standards.[9] However, implementation remains ostensibly contested and uneven.

Of course, beyond the rules, there are social and cultural norms that influence public interpretations about the appropriateness and acceptableness of online speech,[10] as well as institutional mechanisms of external oversight (such as Meta's Oversight Board) and evolving approaches to content moderation that is increasingly moving away from established paradigms like professional fact-checking toward more participatory models, including community notes[11]. These governance layers intersect in complex ways, creating spaces where normal policy responses prove insufficient and exceptional decision-making often becomes necessary, whether through additional content policies, contextual exceptions, manual review of escalated content, or specialized institutional arrangements.

What is problematic in its obviousness, moreover, is that the processes of rule- and norm-setting are developed within a limited set of geopolitical and cultural reference points, effectively extending privately curated rule systems across platforms' global user bases. By contrast, actors in Majority World contexts are rarely substantively or meaningfully engaged in these processes in ways that reflect their lived realities or normative frameworks, resulting in the operationalization of homogenized governance regimes across more than 85% of the world's population without adequate mechanisms for contextual adaptation or epistemic inclusion. Understanding how platforms navigate these ambiguous areas requires examining not only what borderline content is, but how it has come to be understood and operationalized within the unequal systems of global platform governance.

---

9 Meta, Facebook's Corporate Human Rights Policy, March 2021, https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy. pdf; TikTok, Upholding Human Rights, accessed February 18, 2026, https://www.tiktok.com/transparency/en/upholding-human-rights; X (formerly Twitter), Defending and Respecting Our Users' Voice, accessed February 18, 2026, https://help.x.com/en/rules-and-policies/defending-and-respecting-our-users-voice.

10 Amélie Heldt, "Borderline Speech: Caught in a Free Speech Limbo?," *Internet Policy Review*, October 15, 2020, https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510.

11 Yuwei Chuai et al., "Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?," *Proc. ACM Hum.-Comput. Interact.* 8, no. CSCW2 (2024): 428:1-428:52, https://doi.org/10.1145/3686967.

TECHGLOBAL INSTITUTE

Borderline content:
a vague umbrella term

For many practitioners as well as researchers, the term "borderline content" is largely "vague and definitionally complex."[12] In fact, borderline content was primarily adopted as an actor language by major platforms rather than a stable analytical category, which reflects both the practical challenges platforms face, and their attempts to manage public expectations about the willingness and capabilities of dealing with gray area cases.[13]

In Mark Zuckerberg's 2018 "Blueprint for Content Governance and Enforcement," he described how users tend to engage with content that approaches but does not cross lines of policy violation.[14] Internal research at Facebook suggested a "natural pattern of borderline content getting more engagement" for materials such as nudity, sexual suggestiveness, and potentially offensive content that falls short of hate speech definitions. Zuckerberg argued that the pattern persists "no matter where we draw the lines,"[15] implying that moderation must change the incentive and reduce the visibility and distribution of such content. In 2021, Meta subsequently published its Content Distribution Guidelines, and later created a dedicated policy section for content that would be considered "borderline to the community standards".[16] Some examples are reproduced in this section (see Table 1), including borderline nudity and sexual activity, violence and graphic, and bullying and harassment materials, whose distributions are meant to be restricted as they might bring down the quality of platform discourse.

12 Stuart Macdonald and Katy Vaughan, "Moderating Borderline Content While Respecting Fundamental Values," *Policy & Internet* 16, no. 2 (2024): 347–61, https://doi.org/10.1002/poi3.376.

13 Tarleton Gillespie, "Do Not Recommend? Reduction as a Form of Content Moderation," *Social Media + Society* 8, no. 3 (2022), https://doi.org/10.1177/20563051221117552.

14 Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement | Facebook," Facebook, November 15, 2018, https://perma.cc/ZK5C-ZTSX. See also Josh Constine, "Facebook Will Change Algorithm to Demote 'Borderline Content' That Almost Violates Policies," *TechCrunch*, November 15, 2018, https://techcrunch.com/2018/11/15/facebook-borderline-content/.

15 Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement | Facebook."

16 Meta, "Content Borderline to the Community Standards | Transparency Centre," *Meta*, accessed July 28, 2025, http://archive.today/2025.07.28-054123/https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards.

Nevertheless, recent reporting quoting Meta whistleblowers suggests that the platform would limit enforcement against certain forms of borderline harmful content when the stock price is down.[17] While such accounts should be interpreted with caution, they point to the broader role of economic and strategic incentives in shaping how the policies are actually enforced in practice when platforms face competition and revenue drop.

Relevant definitions first appeared in Google's steps taken against online terrorism, where the platform announced that for certain content that do not clearly violate its content policies, such as inflammatory religious or supremacist content, YouTube would add warning labels, de-monetize, reduce from recommendations, and make them not eligible for comments or user endorsements.[18] Later in January 2019, YouTube announced that it would begin reducing recommendations for borderline content that "comes close to — but doesn't quite cross the line of — violating [their] Community Guidelines."[19] While YouTube's community guidelines seek to explicitly define what constitutes borderline content, the definition is circular and stipulative, that is, it explains the term using the term itself without clarifying its substance, leaving it both operationally and conceptually vague for users and moderators alike.

That being said, the platform's approach becomes somewhat understandable when examined alongside its search quality evaluation framework. In late 2019, YouTube detailed how the platform relies on external evaluators (distinct from content moderators) worldwide to assess video quality, with their consensus input used to train machine learning systems.[20] Notably, all third-party evaluators are trained using publicly accessible guidelines that reveal the platform's quality assessment criteria.[21] A deep dive into these search quality rating guidelines (last updated in January 2025) uncovers the considerations that inform borderline content decisions, as many criteria for evaluating search results apply similarly to user-generated content. The guidelines instruct evaluators to consider factors such as the purpose of the content, its potential to cause harm, the reputation and expertise of creators, and overall trustworthiness. These

17 Marianna Spring and Mike Radford, "Meta and TikTok Let Harmful Content Rise after Evidence Outrage Drove Engagement - Whistleblowers," BBC News, March 16, 2026, https://www.bbc.co.uk/news/articles/cqj9kgxqjwjo.

18 Kent Walker, "Four Steps We're Taking Today to Fight Terrorism Online," Google, June 18, 2017, https://blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-ter ror/.

19 The YouTube Team, "Continuing Our Work to Improve Recommendations on YouTube," YouTube Official Blog, January 25, 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve/.

20 YouTube, "The Four Rs of Responsibility, Part 2: Raising Authoritative Content and Reducing Borderline Content and Harmful Misinformation," YouTube Official Blog, March 12, 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/.

21 "Introduction to Search Quality Rating," January 23, 2025, https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf?sjid=5405440185927561652-EU#page=58.04.

criteria operationalize what YouTube considers "borderline" — content that may not violate specific community guidelines but might score poorly on quality dimensions like credibility, creator authority, or potential for harm.

While not explicitly phrased as such, other platforms have developed parallel categorization and moderation schemes. For instance, TikTok launched "For You Feed Eligibility" standards that describe content "made ineligible for the For You Feed when they may not be suitable for a broad audience,"[22] seemingly the same approach under different policy phrasing. Before its acquisition by Elon Musk, Twitter used to maintain "sensitive media policy" for material that, while not violating rules, may "depict sensitive topics," including graphic violence and consensual sexual content.[23] After Musk's take-over, however, X discarded this policy stream and instead leaned on expanded adult content as well as violent content policies.[24] That said, in terms of policy stance, most platforms appear to choose to reduce the circulation and visibility of borderline content, as a way of striking the balance between protecting users' freedom of expression while preventing potential harm.

Platform	Meta	YouTube	TikTok	Twitter (before Musk)	LinkedIn
Specific mention of "borderline content"	Yes	Yes	No	No	No
Relevant definitions	Categories developed specifically to capture types of content that are not prohibited by its Community Standards, but that come close to the lines drawn by those policies.	Content that comes close to — but does not quite cross the line of — violating its Community Guidelines.	Content is made ineligible for the For You Feed when they may not be suitable for a broad audience.	Media that depict sensitive topics, including violent and adult content.	-

22 TikTok, "For You Feed Eligibility Standards," TikTok, March 8, 2023, https://www.tiktok.com/community-guidelines/en/fyf-standards/.

23 Twitter, "Sensitive Media Policy," Twitter Help Center, January 2022, https://web.archive.org/web/20220701010743/https://help.twitter.com/en/rules-and-policies/media-policy.

24 X Help Center, "Adult Content Policy," May 2024, https://help.x.com/en/rules-and-policies/adult-content.

Platform	Meta	YouTube	TikTok	Twitter (before Musk)	LinkedIn
Specific content categories subject to these content policies	- borderline adult nudity and sexual activity - borderline bullying and harassment, hate speech, and violence and incitement (e.g., content that dehumanizes individuals or groups who are not defined by their protected characteristics, praises or mocks past violence and harm, or uses profanity targeted towards a group) - content that does not violate the COVID-19 or vaccine policies, but which shares misleading or sensationalized information about vaccines in a way that would be likely to discourage vaccinations	- videos that contain inflammatory religious or supremacist content - content that might misinform users in a harmful way	showing non-explicit imagery of youth sexual abuse in fiction or art, such as television shows with scenes of implied sexual acts	- graphic violence - consensually produced adult content	-
Actions taken	Reduce the distribution	Reducing recommendations, demonetization, add warning labels, not eligible for comments or user endorsement	Not allowed in the For You Feed	Sensitive account feature, prohibited from be included in live video, profile, header, or banner images	-

Table 1: Relevant definitions, content categories, and actions across major platforms for borderline cases. Policies accessed on August 19, 2025.[25]

25 Meta, "Content Borderline to the Community Standards | Transparency Centre"; YouTube, "The Four Rs of Responsibility, Part 2"; Twitter, "Sensitive Media Policy"; TikTok, "For You Feed Eligibility Standards."

Beyond industry, the term has also increasingly gained prominence in policy discussions, particularly about borderline terrorist and violent extremism content, convened by organizations such as the GIFCT and the Christchurch Call to Action.[26] Arguably, the absence of clear definitional parameters also complicates efforts to develop consistent and transparent moderation policies across different platforms and jurisdictions.[27]

Academic and practitioner communities have developed two primary paradigms for understanding these hard-to-adjudicate cases. First, the *legal paradigm* primarily frames these challenges through the concept of **"lawful but awful"** content, referring to content that, while not explicitly illegal, might still promote extremist narratives, incite violence, or generate significant social harm.[28] Legal scholars focus particularly on content that exists at the boundaries of protected speech categories (e.g., hate speech in the US context), where cultural, political, or jurisdictional differences create interpretive uncertainty. Second, the *trust and safety paradigm* approaches the same phenomena from an operational perspective, focusing on the practical challenges of scale, consistency, and institutional design that platforms face when implementing content policies. As one trust and safety expert explained:

> **"The area of work is really defined as being these kinds of very high risk cases… where the typical policy response wouldn't be appropriate for managing that type of risk. … we wouldn't necessarily be calling it 'borderline content,' but… ones where there may need to be a different type of consideration with how we're responding." (ID2)**

26 Erin Saltman and Micalie Hunt, *Borderline Content: Understanding the Gray Zone* (Global Internet Forum to Counter Terrorism, 2023), https://gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf; ICCT, "Feasibility Study On The Development Of An Assessment Framework For Terrorist And Borderline Content Online," International Centre for Counter-Terrorism - ICCT, accessed July 30, 2025, https://icct.nl/project/feasibility-study-development-assessment-framework-terrorist-and-borderline-content-online.

27 Ellie Rogers, "The Need for Greater Transparency in the Moderation of Borderline Terrorist and Violent Extremist Content," *Internet Policy Review* 14, no. 3 (2025), https://doi.org/https://doi.org/10.14763/2025.3.2012.

28 Daphne Keller, "Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users," *The University of Chicago Law Review Online,* June 28, 2022, https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech/.

Content -
What are we looking at?

The content dimension focuses on observable characteristics of material under review, including the modality of the content (text, image, video, or audio), the topics it addresses (such as religious, political, or sexual themes), whether it is fictional or real, and visual elements it features and the linguistic features it employs.

Most platforms have increasingly systematized these considerations through frameworks. For example, YouTube's YMYL (Your Money or Your Life) topic policy, which identifies content related to health, financial security, or public institutions where "the topic itself is harmful or dangerous" or could "cause harm if the content is not accurate and trustworthy."[29] TikTok and Twitter/X also identified sensitive themes related in particular to graphic violence and sexual suggestiveness content. Borderline cases often arise when these features produce conflicting interpretive signals, particularly in multi-modal content. A single piece of content may combine benign visuals, coded audio, and culturally specific language, each pointing toward different moderation outcomes. In such cases, determining which modality should be prioritized and how to weigh competing meanings becomes inherently challenging for moderators (ID8).

A case raised by a child safety expert (ID5) illustrates this ambiguity. Bart Simpson's naked skateboarding scenes from the 2007 movie, long treated as comedic, non-sexualized elements of a mainstream animated series, offers a useful entry point into the complexities in content moderation. Animated or artistic representations that would clearly violate policies if depicting real people create genuine borderline cases when the "harm" becomes more abstract. Should nudity policies designed to protect children and women apply to cartoon characters like Bart Simpson, whose representation is mediated through satire, genre conventions, and audience expectations? Even if one answers this question in the affirmative, would the assessment remain the same if, instead of an established production house, a user generates or modifies nude cartoons of Bart Simpson using image-generating or editing tools, or if a comparable level of nudity is depicted for Marge, Lisa, or Maggie Simpson?

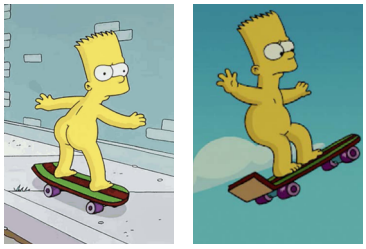


Figure 2: Bart Simpson skateboarding naked in Simpsons Movie (2007).

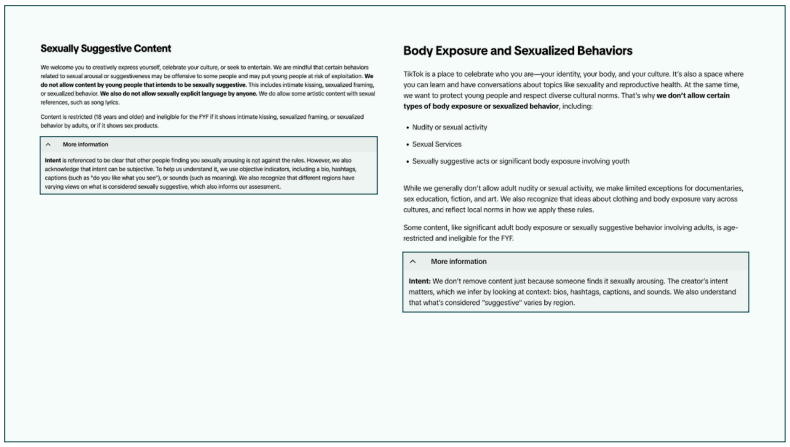
29 "Introduction to Search Quality Rating."

Conduct -

Why was it posted, and by whom?

Beyond content characteristics, platforms increasingly turn to conduct-based assessment — examining user intent, creator reputation and history, posting patterns, and broader behavioral contexts. This shift from "what was posted" to "why was it posted" represents a move to more proactive interventions, but also introduces new forms of interpretive complexity.

TikTok's evolving approach to sensitive and mature themes illustrates this shift. Earlier versions of its policies included a distinct category for "sexually suggestive content" focusing on behaviors related to sexual arousal or suggestiveness that might be considered inappropriate or offensive. In a subsequent update effective September 13, 2025, TikTok restructured these categories while retaining and further emphasizing its approach to inferring user intent in cases involving potentially suggestive content. Under this model, TikTok evaluates so-called "objective indicators" of intent, including user bios, hashtags, captions (e.g., suggestive phrasing), and audio elements. In this sense, behavioral inference is derived from patterns of use rather than the content in isolation or its immediate perception by viewers.



Sexually Suggestive Content

We welcome you to creatively express yourself, celebrate your culture, or seek to entertain. We are mindful that certain behaviors related to sexual arousal or suggestiveness may be offensive to some people and may put young people at risk of exploitation. **We do not allow content by young people that intends to be sexually suggestive.** This includes intimate kissing, sexualized framing, or sexualized behavior. **We also do not allow sexually explicit language by anyone.** We do allow some artistic content with sexual references, such as song lyrics.

Content is restricted (18 years and older) and ineligible for the FYF if it shows intimate kissing, sexualized framing, or sexualized behavior by adults, or if it shows sex products.

More information

Intent is referenced to be clear that other people finding you sexually arousing is not against the rules. However, we also acknowledge that intent can be subjective. To help us understand it, we use objective indicators, including a bio, hashtags, captions (such as "do you like what you see"), or sounds (such as moaning). We also recognize that different regions have varying views on what is considered sexually suggestive, which also informs our assessment.

Body Exposure and Sexualized Behaviors

TikTok is a place to celebrate who you are—your identity, your body, and your culture. It's also a space where you can learn and have conversations about topics like sexuality and reproductive health. At the same time, we want to protect young people and respect diverse cultural norms. That's why **we don't allow certain types of body exposure or sexualized behavior**, including:

- Nudity or sexual activity
- Sexual Services
- Sexually suggestive acts or significant body exposure involving youth

While we generally don't allow adult nudity or sexual activity, we make limited exceptions for documentaries, sex education, fiction, and art. We also recognize that ideas about clothing and body exposure vary across cultures, and reflect local norms in how we apply these rules.

Some content, like significant adult body exposure or sexually suggestive behavior involving adults, is age-restricted and ineligible for the FYF.

More information

Intent: We don't remove content just because someone finds it sexually arousing. The creator's intent matters, which we infer by looking at context: bios, hashtags, captions, and sounds. We also understand that what's considered "suggestive" varies by region.

Figure 3: Evolution of TikTok's policies on sensitive and mature themes and the role of intent in moderation. The left panel shows the version of the policy eff ective from May 17, 2024 to September 13, 2025, which included a distinct category for "sexually suggestive content."[30] The right panel shows the updated guidelines eff ective September 13, 2025,[31] in which content categories were consolidated and greater emphasis was placed on inferring user intent through contextual and behavioral indicators.

30 TikTok, "Sensitive and Mature Themes," TikTok, April 17, 2024, https://web.archive.org/web/20250817061810/https://www.tiktok.com/community-guidelines/en/sensitive-mature-themes.

31 TikTok, "Sensitive and Mature Themes," TikTok, August 14, 2025, https://www.tiktok.com/community-guidelines/en/sensitive-mature-themes.

By incorporating conduct-based signals, platforms are better able to identify forms of harm that may not be evident from content alone. At the same time, reliance on inferred intent expands moderation into more subjective terrain, where cultural interpretation, linguistic nuance, and platform-specific norms play a decisive role. More critically, these behavioral signals often embed implicit cultural assumptions that can systematically disadvantage certain user populations. A striking example arises in financial fraud detection, where platforms have used rooted Android devices as a proxy for suspicious activity. As one behavioral integrity expert noted, "rooted devices are actually very common in Latin America and parts of Asia, and treating them as a signal of fraud would be misleading" (ID6).

A case decided by Meta's Oversight Board further illustrates how conduct-based signals shape content moderation outcomes involving borderline content targeting specific demographic groups. In this case,32 which involved the use of monkey emojis to refer to Black people, the board emphasized that meaning emerged not from the emoji in isolation but from patterns of usage, contextual cues, and coordinated behavior, underscoring that "algospeak" requires platforms to assess intent, audience targeting, and behavioral context rather than rely solely on surface-level content. The decision highlighted that failures to adequately interpret conduct signals across both automated and human moderation systems, such as repeated emoji use during sensitive moments like major sporting events, can allow coded racist and discriminatory campaigns and the dehumanization of individuals based on protected characteristics to persist. These concerns are not merely theoretical: Meta's automated systems have previously misclassified content, including a 2021 incident in which a recommendation feature labeled Black men as "primates," prompting the feature's removal.[33] In the case under discussion, the Oversight Board recommended that Meta strengthen its detection of violative emoji and other "algospeak" through regular audits of training data and ensure that enforcement processes consistently route content to reviewers with appropriate linguistic, cultural, and regional expertise.

33 AFP, "Facebook Mistakenly Labels Black Men 'Primates,'" Al Jazeera, September 4, 2021, https://www.aljazeera.com/news/2021/9/4/facebook-mistakenly-labels-black-men-primates.

TECHGLOBAL INSTITUTE

Context -
What frameworks apply?

The contextual dimension encompasses the multiple, often conflicting frameworks that determine how identical content and conduct are interpreted: situational factors (newsworthiness, crises, educational value), legal jurisdictions, cultural sensitivities, and institutional considerations. Context matters the most in determining borderline cases, which reveals the power dynamics and biases embedded in platformized content governance.

Contextual exceptions illustrate how platforms attempt to balance competing values. Meta's newsworthiness allowance[34] and YouTube's Educational, Documentary, Scientific and Artistic (EDSA) framework,[35] for example, allows content that would otherwise violate guidelines when it serves certain legitimate purposes. TikTok's recently added "public interest exceptions" similarly protect educational, documentary, medical, scientific, artistic, counterspeech, and satirical content.[36] These frameworks recognize that identical content characteristics and user conduct can have entirely different meanings depending on context.

For instance, a policy advisory opinion by Meta's Oversight Board illustrates how the absence of contextual sensitivity can render content moderation "overbroad, and substantially and disproportionately restricts free expression."[37] Meta imposed a blanket ban on the term "shaheed," relying on a linear interpretation that treated the word as inherently constituting praise for terrorists or other dangerous individuals, thereby collapsing a linguistically and culturally complex term into a single, punitive meaning. By disregarding the term's multiple uses across languages, religions, and sociopolitical contexts,

34 Meta, "Approach to Newsworthy Content," November 12, 2024, https://transparency.meta.com/en-gb/features/approach-to-newsworthy-content/.

35 YouTube Help, "How YouTube Evaluates Educational, Documentary, Scientific & Artistic (EDSA) Content," YouTube Help, accessed July 30, 2025, https://support.google.com/youtube/answer/6345162?sjid=6362539677884254510-EU.

36 TikTok, "Enforcement," TikTok, August 14, 2025, https://www.tiktok.com/community-guidelines/en/enforcement?cgversion=2025H2update#1.

37 Oversight Board, *Meta Accepts Key Oversight Board Recommendations to End Blanket Ban on "Shaheed,"* July 2, 2024, https://www.oversightboard.com/news/meta-accepts-key-oversight-board-recommendations-to-end-blanket-ban-on-shaheed/.

ranging from personal names and neutral description to mourning, honorific usage, or historical reference, Meta's approach disproportionately restricted lawful expression among both Arabic- and non-Arabic-speaking users across countries in the Global South. It emphasized that while preventing incitement to violence is a legitimate and necessary objective, context-blind moderation anchored in keyword-based presumptions risks suppressing speech that neither glorifies violence nor supports extremist causes, ultimately undermines freedom of expression, alienates affected communities, and reinforces structural bias and power asymmetries embedded in platformized speech governance.

A more recent decision by the Oversight Board, involving Kenyan political discourse and the term "tugeges," overturned Meta's removal decision, finding that the platform had over-relied on a static conduct signal (its prior classification as a slur) without sufficient attention to evolving linguistic context and local political usage.[38] It emphasized that the meaning and impact of the term were temporally and contextually contingent, noting that while it may have carried stronger ethnic connotations during the 2022 Kenyan elections, its usage had since shifted toward broader political critique, underscoring the need for context-sensitive designation processes and sustained engagement with local stakeholders when defining and enforcing speech restrictions.

Another illustrative example concerns the cultural significance attached to bare-breasted women in the Global South, which reveals how content moderation regimes often misinterpret non-Western forms of embodiment and expression. For many Indigenous communities worldwide — ranging from the Himba in Namibia[39] and the Zulu People in South Africa[40] to the Yanomami in Brazil and Venezuela,[41] the Asmat[42] and Korowai[43] Peoples in Indonesia, and the Jarawa[44] and Sentinelese[45] Peoples in India, among numerous others across Africa,

38 Oversight Board, *Comment on Kenyan Politics Using a Designated Slur | Oversight Board*, December 9, 2025, https://www.oversightboard.com/decision/fb-o65b8pn7/; "The 'Tugeges' Test: How a Kenyan Word Exposed the Fault Lines In," accessed March 16, 2026, https://www.unesco.org/en/articles/tugeges-test-how-kenyan-word-exposed-fault-lines-global-content-moderation.

39 "Women of the Himba Tribe," Google Arts & Culture, accessed March 27, 2026, https://artsandculture.google.com/story/women-of-the-himba-tribe/DwWRp9f7QXLkCQ.

40 Siyabona Africa, "Zulu - African Culture," accessed March 27, 2026, https://www.krugerpark.co.za/africa_zulu.html.

41 Survival International, "Yanomami," accessed March 27, 2026, https://www.survivalinternational.org/peoples/yanomami.

42 "Unveiling the Hidden World of the Asmat Tribe in West Papua," Aqua Expeditions, accessed March 27, 2026, https://www.aquaexpeditions.com/blog/asmat-tribe-west-papua.

43 "Facts About Korowai Tribe in Southern Papua | Authentic Indonesia Blog," *Authentic Indonesia*, n.d., accessed March 27, 2026, https://authentic-indonesia.com/blog/facts-about-korowai-tribe-in-southern-papua/.

44 Survival International, "Ang," accessed March 27, 2026, https://survivalinternational.org/peoples/ang.

45 Survival International, "Sentinelese," accessed March 27, 2026, https://survivalinternational.org/peoples/sentinelese.

the Americas, Asia, and Australia — full or partial nudity constitutes an intrinsic and normalized aspect of social life, tradition, and cultural identity, rather than a marker of sexual provocation. Practices such as ochre painting, scarification, tattoos, jewelry, and other bodily adornment function as nonverbal expressions of personhood, social roles, cosmological beliefs, harmony with the environment, spiritual worldviews, and collective heritage. When situated within these cultural frameworks, dominant assumptions that equate exposed female breasts with immodesty, moral transgression, or sexualization are revealed to be culturally contingent rather than universal.[46] Yet platform moderation practices have repeatedly failed to account for such contextual meanings. Meta, for instance, has a documented history of removing content depicting bare-breasted Indigenous women in ceremonial, historical, or educational settings (see Figure 4), while simultaneously permitting sexualized representations of women that conform to Western norms, such as micro-bikinis or commercialized displays of nudity, as well as other forms of nudity deemed acceptable under platform policies, including breastfeeding and exposure resulting from medical procedures such as mastectomy.

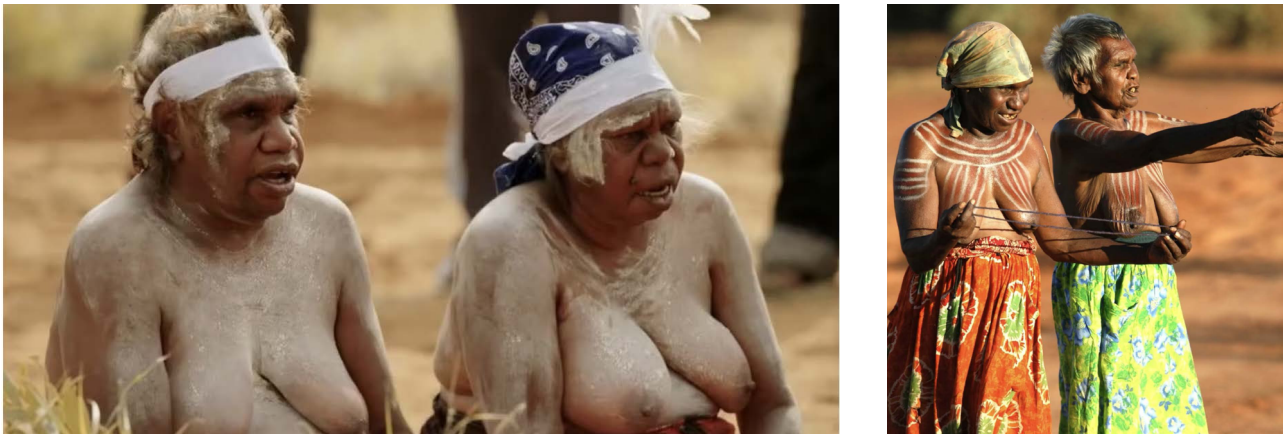


Figure 4: Facebook removed a 2:46-minute television trailer in 2015 that featured two scenes of elderly Aboriginal women having their bare breasts painted with ochre as part of a traditional ceremony (left),[47] and in 2016 suspended user profi les for sharing an article on Aboriginal feminism accompanied by images of two bare-breasted Indigenous women in traditional attire (right).[48]

46 Tech Global Institute, *PUBLIC COMMENTS ON CASES 2025-005-IG-UA, 2025-006-IG-UA, 2025-007-IG-MR, AND 2025-008-FB-MR ASSESSING IMPACT OF META'S NUDITY RULES ON INDIGENOUS PEOPLES*, June 2025, https://techglobalinstitute.com/wp-content/uploads/2025/06/Comment-on-the-Impact-of-Metas-Nudity-Rules-on-Indigenous-Peoples-Tech-Global-Institute-2.pdf.

47 Emilia Terzon, "Aboriginal Video Pulled Due to Facebook's Nudity Guidelines," ABC News, April 13, 2015, https://www.abc.net.au/news/2015-04-13/indigenous-video-pulled-facebook-nudity-rules/6388090.

48 Penny Timms, "Facebook Suspends Users over 'nude' Photo of Aboriginal Women in Traditional Dress," ABC News, March 14, 2016, https://www.abc.net.au/news/2016-03-14/facebook-suspends-users-over-nude-photo-of-aboriginal-women/7244710.

Since then, inconsistent approaches to such borderline content have led to multiple cases being escalated to Meta's Oversight Board,[49] which has observed that both blanket prohibitions on images of bare-chested Indigenous women and ad hoc exceptions fail to meet standards of necessity and proportionality; such approaches disproportionately restrict Indigenous women's rights to self-expression, cultural representation, and information sharing, while reinforcing their exclusion from digital spaces. It emphasized that such cultural expression should be afforded the same contextual consideration already extended to other forms of nudity permitted under existing exceptions, such as bare-breasted acts of protest, and has urged Meta to develop clearer escalation pathways and internal guidelines with clearly articulated contextual indicators (including visual cues such as body paint, jewelry, feathers, cultural symbols, or ceremonial objects) to support more consistent and culturally attuned moderation at scale. However, this exception "should be applied on escalation only … where it reflects socially accepted custom and belief, and does not misrepresent these practices."

Legal frameworks add another layer of complexity, particularly when local laws in certain countries conflict with international human rights standards. The ongoing challenge of blasphemy content in Pakistan illustrates this tension. Authorities frequently issue takedown requests targeting religious minorities like the Ahmadiyya community.[50] As a minority sect that is constitutionally designated as non-Muslims and that faces systemic religious persecution and discrimination in Pakistan, content related to the Ahmadiyya community may technically be unlawful within the jurisdiction. However, as an expert legal counsel explained: "Even if the content is clearly unlawful within the jurisdiction, the human rights team took a clear stance that if they block it, the company will be in clear violation of Article 19 of the ICCPR and the company's GNI commitments" (ID1).

Similarly, in Bangladesh, following the ouster of the Awami League in the wake of the 2024 Monsoon Revolution and its subsequent designation as a terrorist organization under amended counterterrorism laws, content expressing support for the party, its leaders, or its political advocates is treated as legally

49 Oversight Board, "Images of Partially Nude Indigenous Women | Oversight Board," *Oversight Board*, June 3, 2025, https://www.oversightboard.com/decision/bun-s8141rab/.

50 Megha Rajagopalan, "Pakistan Forced Down Apps Made By A Persecuted Religious Minority," BuzzFeed News, February 4, 2021, https://www.buzzfeednews.com/article/meghara/pakistan-forced-down-ahmadiyya-apps.

proscribed, even where such expression constitutes ordinary political speech unrelated to terrorism or incitement to violence.[51] This reclassification effectively collapses contextual distinctions between political affiliation and violent extremism, creating acute challenges for platforms attempting to reconcile domestic legal compliance with international human rights standards on freedom of expression. Moreover, linguistic features create persistent borderline challenges. As an Indonesian moderator explained, regional slang from Bali might be entirely innocent in the specific local context, but could refer to something that might be considered as sexually suggestive when encountered by users from the other regions of Indonesia (ID8). The same textual content literally transforms meaning as it moves across cultural boundaries, illustrating why content-focused analysis alone proves insufficient.

In practice, these decisions around these borderline cases require platforms to develop sophisticated expertise and capacity. Moreover, it highlights why such decisions may appear arbitrary to affected communities who lack visibility into the full analytical process. Who, then, is responsible for making these complex assessments, and how do platforms organize the human and technological infrastructure required to navigate such interpretive complexity at the scale of billions of posts per day? Understanding the governance of borderline cases requires mapping the decision-making ecosystem that transforms the moderation principles into concrete outcomes.

51 Human Rights Watch, *Bangladesh: New Crackdown Under Anti-Terrorism Law | Human Rights Watch*, October 8, 2025, https://www.hrw.org/news/2025/10/08/bangladesh-new-crackdown-under-anti-terrorism-law; TBS Report, "Awami League, All Its Affiliates Now Officially Banned," The Business Standard, May 12, 2025, https://www.tbsnews.net/bangladesh/politics/awami-league-banned-1140321.

The hidden machinery of borderline content moderation

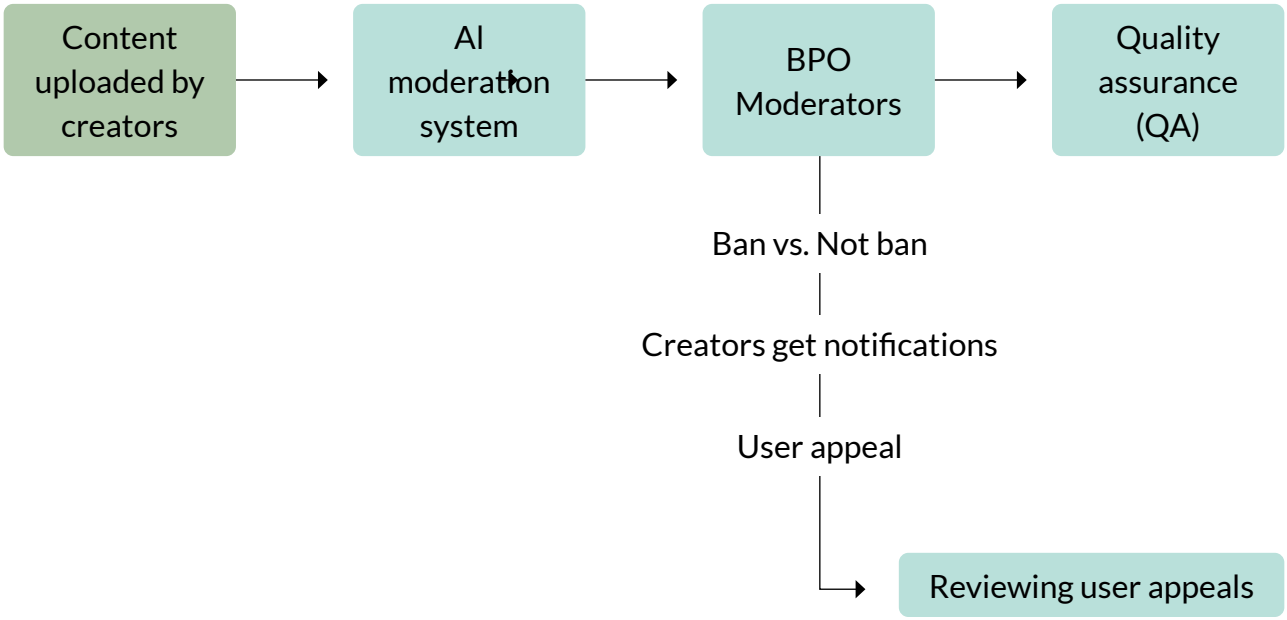


Figure 5.1: Approximate configuration of actors engaged in content moderation decision-making, based on interviews with multiple trust & safety professionals across platforms.

As shown in Figure 5, for most major user-generated content platforms, content uploaded by creators typically passes through a layered moderation pipeline. They are first processed by AI moderation systems, designed to handle the sheer volume of material at scale.[52] However, these automated systems have proved insufficient when it comes to borderline cases, which are typically flagged for human review.

Human moderation is often distributed across a hybrid ecosystem. While some review functions are handled internally by platform policy teams, a significant

52 Robert Gorwa et al., "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance," *Big Data & Society* 7, no. 1 (2020): 1–15, https://doi.org/10.1177/2053951719897945.

proportion is outsourced to third-party vendors, including business process outsourcing (BPO) firms, law firms, and specialized moderation contractors. Within these environments, moderators develop informal practices to manage ambiguity.[53] One such practice is the use of what practitioners describe as "market alignment trix" (ID7) — a list of agreements written informally (with or without clients' knowledge) between content moderators and the quality assurance (QA) team to refer to to avoid mismatch between them, in a time where the content moderation policy is not sufficient to help the moderators make judgments on grey area cases. Describing the daily rhythm of this coordination, another interviewee mentioned that usually if the morning shift already decides which policy they would use, the rest will follow on that (ID8). The precedent-setting decisions made during one shift influence subsequent determinations, creating informal consistency mechanisms that operate alongside formal policy guidance. Yet these internal alignment practices do not create binding consensus or precedent beyond the firm or individual platform. Because such interpretive heuristics are developed and applied within platform-specific governance structures, each firm adopts divergent policies and thresholds for similar borderline cases, resulting in inconsistent moderation outcomes both within and across platforms.

QA teams function as crucial intermediaries in this ecosystem, reviewing flagged cases from front-line moderators and channeling feedback to platform clients while providing local expertise and cultural competency. The QA layer, positioned between external BPO moderators and internal platform content enforcement teams in the diagram, serves as both quality control and cultural translation mechanism. Consider how this plays out with LGBTQ+ content moderation in Indonesia, where cultural and religious sensitivities require delicate navigation between "presentation" and "glorification" of LGBTQ+ groups. Platforms normally would leave the policy vague without defining precisely what constitutes glorification, and the QAs would need to play the words with platforms to allow a more localized adaptation within the moderator team (ID8).

53 *Trust & Safety Market Research Report (Duco, 2024), https://duco-public-static-assets.s3.amazonaws.com/Duco+TnS+MRR-+FINAL.pdf.*

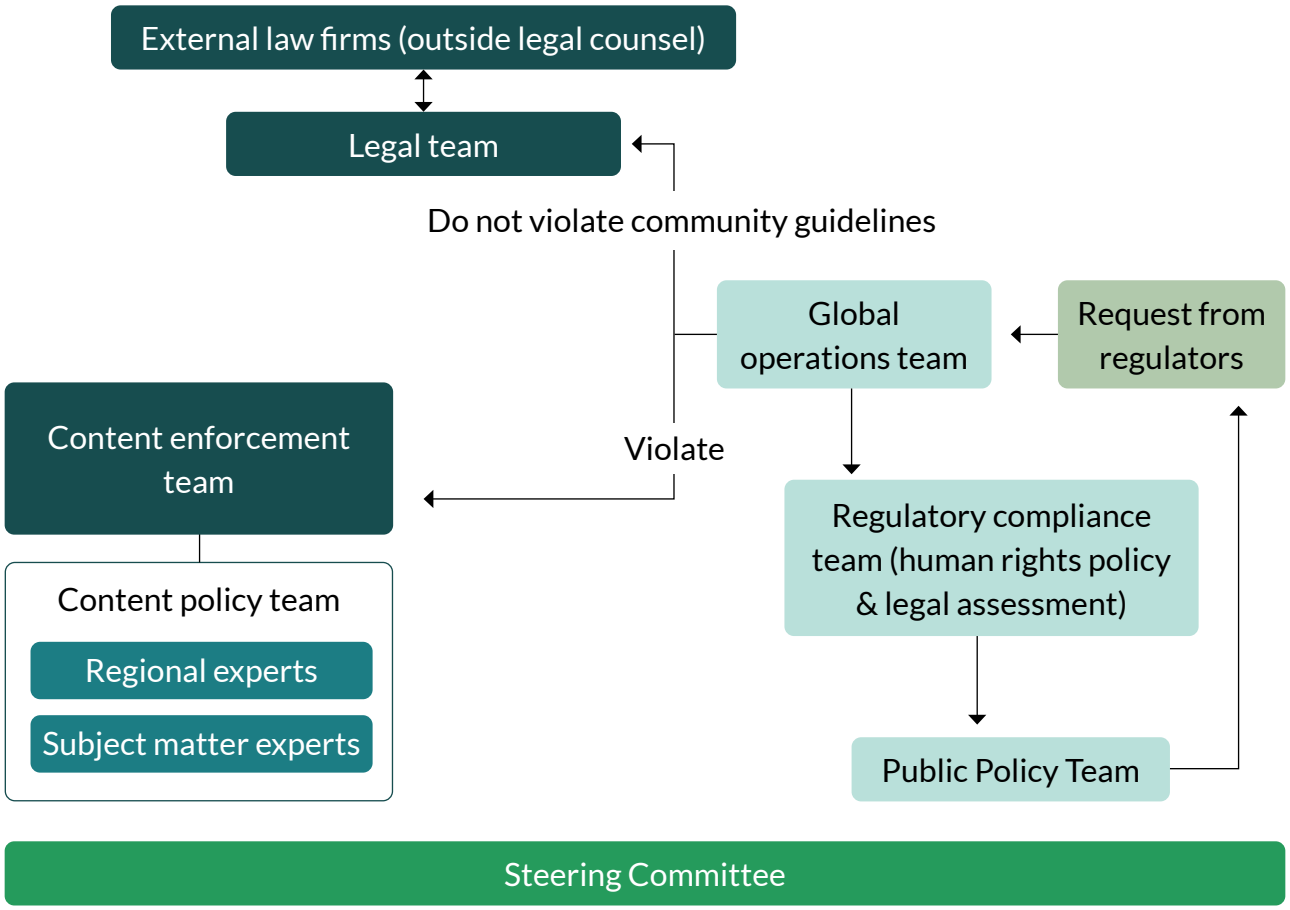


Figure 5.2: Approximate configuration of actors engaged in content moderation decision-making, based on interviews with multiple trust & safety professionals across platforms.

As cases move beyond vendor-managed initial review, they enter the platform-side governance apparatus shown in the upper portions of the diagram. Here, different organizational units — content enforcement teams, legal departments, global operations, regulatory compliance teams, and public policy groups — intersect in complex ways that reflect both functional specialization and organizational hierarchy. Some platforms maintain dedicated subject matter experts and regional specialists, with the extent of organizational localization varying significantly across companies. For particularly complex cases that resist standard escalation processes, some specialized institutional arrangements, such as Meta's Oversight Board or Google's Steering Committee come into play. When governments issue takedown requests or when new regulations or political crises create coercive compliance requirements, the external pressure would also influence the context evaluation of the moderation cycle.

Challenges in everyday decision-making around borderline content

The governance architecture described above might appear sophisticated on paper, but its everyday operation reveals profound systemic challenges that extend far beyond individual content decisions. These challenges expose fundamental questions about power, representation, and equity in global content governance, creating patterns of marginalization that affect billions of users while remaining largely invisible to public scrutiny.

Centralized decision-making and the systematic marginalization of local cultural expertise

Perhaps the most striking feature of current borderline content governance architecture across most major user-generated content platforms examined in this study is the systematic concentration of decision-making authority in Western, English-speaking corporate headquarters, even as the majority of both content creation and content moderation work occurs in Majority World countries. This geographic distribution of power creates what practitioners describe as fundamental disconnects between those who make and implement policies and those who are subject to their enforcement. Several interviewees feel that within platforms, executives who really make the decisions are in the United States, while everyone else is consulted. This dynamic plays out in countless individual cases where cultural expertise exists but lacks enforcement authority. One trust and safety expert recalled a case involving Swahili content where a junior employee's local knowledge contradicted executive judgment, and the final moderation decision ultimately deferred to executive authority, with senior leadership's interpretation taking precedence over contextually grounded expertise (ID2). Another interviewee

observed that the organizational hierarchy reinforces these inequities systematically:

> **"You look at the trust and safety organizational hierarchy, the most senior people who clearly have veto ability … are all sitting either in Western Europe or in the United States. In the Global Majority regions, you have reviewers, you have line managers, you have some country representatives… but ultimately the escalation path leads to Western Europe or the United States." (ID3).**

The centralized, hierarchical decision-making underlies a significant mismatch between those who have the local expertise and cultural competency, which create huge potential for bias in decision-making around borderline content (ID3). This assessment is corroborated by the cases examined in this study, which demonstrate that erroneous moderation decisions involving borderline content are, in many instances, at least in part attributable to the marginalization of cultural and contextual knowledge; this, in turn, amplifies the potential for bias in content governance outcomes.

The distribution of resources for content moderation reveals another dimension of systemic inequality, creating what might be termed a hierarchy of content governance priorities. Well-resourced regions receive sophisticated moderation capabilities and nuanced policy deliberation, while others struggle with basic content governance infrastructure. A researcher working on content moderation in low-resource languages captured this disparity: "It seems to me that at this very moment in many Majority World countries, the problem is that they don't even have enough resources to do the baseline moderation at all. Borderline content is a very luxurious kind of conversation that Global North is enjoying because they at least have better moderation compared to Global South" (ID4).

This resource gap affects not only moderation quality, but also which borderline cases receive attention. Sophisticated deliberation about

interpretive complexity becomes available primarily in contexts and languages that already possess relatively effective baseline moderation, primarily in English-speaking contexts, concentrating the most nuanced approaches to content governance in regions that may need them least. The problem extends to human resource allocation, where platforms routinely confuse linguistic and geographic expertise. Moderators from one country might be assigned to oversee entire regional operations, despite lacking familiarity with local contexts.[54]

Labor conditions and human costs

The vendor-based model that dominates content moderation creates systematic stress for workers navigating the inherent uncertainties of borderline cases. This tension between the interpretive nature of borderline content and the productivity metrics governing vendor relationships forces front-line moderators to navigate cultural complexity while meeting quantitative performance targets. Human moderators feel that platforms' moderation guidelines are written in a vague way, highlighting how policy ambiguity, perhaps necessary for maintaining flexibility, translates into individual worker stress when combined with performance metrics. The situation worsens as trust and safety markets mature and AI adoption raises productivity expectations, creating what workers describe as exploitation of productivity among the vendors. (ID7). Perhaps most troublingly, moderators report being "punished for subjective judgements" on borderline cases, exactly the kind of nuanced cultural interpretation these cases require. This creates perverse incentives where workers avoid the careful deliberation that effective borderline content governance demands, instead erring on the side of caution and defaulting to mechanically "safe" decisions that shift responsibility away from judgment and onto policy cover, resulting in either over-enforcement under policy allowance or under-enforcement in deference to platform free speech first principles.

The organization of moderation work also limits collective action and peer support. Many workers operate under strict non-disclosure agreements and in fragmented work environments, including remote or home-based settings, where they lack awareness of colleagues performing similar roles. This isolation weakens opportunities for collective organizing and reinforces

54 Mona Elswah et al., *Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages* (Center for Democracy & Technology, 2025), https://cdt.org/insights/content-moderation-in-the-global-south-a-comparative-study-offour-low-resource-languages/.

asymmetries between workers and platform clients (ID10).

Further, these labor dynamics are shaped by regional cultural, economic, and social factors. In South and Southeast Asia, for example, moderation work is not always framed solely as exploitation but also as a source of economic opportunity and social mobility, particularly in contexts where "tech work" carries prestige and stability (ID10). At the same time, this perception coexists with high precarity, limited worker protections, and minimal avenues for collective organization[55]. Gender dynamics further shape the distribution of moderation labor. In some contexts, tasks involving sexually explicit content are informally allocated to male moderators, reflecting cultural norms around propriety and protection, rather than formal platform policies. At the same time, women are overrepresented in certain forms of data work, particularly home-based or flexible arrangements, where such labor is framed as accessible "supplementary income". (ID10)

Technical infrastructure lock-in and cultural responsiveness

The technical infrastructure underpinning borderline content moderation both reflects and entrenches structural biases in how global content governance is designed. Much of today's trust and safety tooling, including detection models, policy taxonomies, annotation pipelines, and risk scoring systems, has been developed within large U.S.-based platforms, primarily for English-speaking or Western user bases. These systems are subsequently productized and exported through vendors and intermediaries to other platforms.

As a result, smaller or regionally focused platforms often inherit moderation architectures that are not designed for their linguistic, cultural, or regulatory contexts. As one interviewee noted, "trust and safety vendors and intermediaries are built around models that worked in Big Tech, but smaller platforms do not rest upon the same type of infrastructure" (ID2). In practice, this means that core components, such as training datasets, classification categories (e.g., hate speech, sexual content), and risk thresholds, are calibrated to Western norms and may fail to capture culturally specific expressions, slang, or context-dependent meanings.

This produces a form of technical lock-in. Platforms become dependent on outsourced tooling, pre-trained models, and vendor-managed workflows that are costly to modify. Building alternative systems

55 Disha Verma, *Labor in the Shadows: Rights and Risks for Asia's Data Workers* (Tech Global Institute, 2025), https://techglobalinstitute.com/wp-content/uploads/2025/08/Labor-in-the-Shadows-TGI.pdf.

like localized datasets, language models, or culturally specific policy frameworks requires significant investment in data collection, annotation, and engineering capacity, which many smaller platforms lack. As a result, they are effectively constrained to adopt moderation systems that reflect the assumptions and priorities of their original developers rather than the diverse needs of global content governance.

Lock-in also operates at the ecosystem level. Shared industry tools, such as hash databases for known harmful content, cross-platform signals, and vendor-provided moderation dashboards, are optimized for standardized categories of clearly violative content (e.g., terrorism or child sexual abuse material). Borderline content, by contrast, is rarely incorporated into these shared infrastructures because it requires contextual and culturally specific interpretation. This limits the ability of platforms to collectively address ambiguous or emerging harms.

The lack of coordination across platforms and jurisdictions further compounds these challenges. While individual companies may develop internal escalation processes, there are few mechanisms for aligning interpretations of borderline content across services or regulatory regimes. Content that is restricted or downranked on one platform may remain fully visible on another, and material that is lawful in one jurisdiction may be contested or harmful in another. Practitioners note that, even where coordination mechanisms exist, "the coordinations within these entangled systems are lacking" (ID5). This fragmentation is particularly consequential for borderline cases, which often only become harmful when viewed across platforms, over time, or within specific cultural contexts. Economic pressures reinforce these dynamics, as platforms increasingly prioritize scalable, generalized moderation solutions, including AI-driven systems, over investments in localized expertise or cross-platform coordination.

Policy recommendations

Taken together, the challenges identified in this research reveal that borderline content moderation is not simply a problem of definitional ambiguity, but of deeper structural constraints. Across content, conduct, and context, ambiguity is compounded by fragmented governance systems, reliance on informal decision-making practices, uneven resource distribution, and technical infrastructures optimized for scale rather than cultural specificity. These dynamics produce inconsistent outcomes, limit the incorporation of local expertise, and systematically disadvantage users and moderators in Majority World contexts.

Addressing these issues therefore requires more than incremental policy refinement. It calls for structural interventions that confront underlying imbalances in power, representation, and capacity within

global content governance. The following recommendations build on practitioner insights and focus on four areas: redistributing decision-making authority, developing contextually responsive governance frameworks, improving labor conditions and professional pathways, and strengthening accountability and cross-platform coordination.

Centering regional expertise

The most critical intervention is to restructure governance so that regional expertise is not merely consulted but empowered. Current systems create what practitioners describe as consultation without authority, where those with deepest cultural and linguistic knowledge function as advisors to distant decision-makers who lack contextual competence to understand their expert advice.

Platforms should develop more decentralized and regionalized decision-making structures with autonomy over policy interpretation and enforcement within their jurisdictions. This includes establishing regional escalation bodies or review panels composed not only of platform employees, but also civil society actors and community representatives with demonstrated cultural competence. Crucially, such authority must be paired with adequate resources — including dedicated budgets, technical infrastructure, and staffing — to ensure these bodies can operate effectively rather than symbolically. A key concern is that decentralization may fragment global policy enforcement and lead to inconsistent standards across jurisdictions, an outcome platforms are likely to resist. However, this shift should not be understood solely as a trade-off between consistency and fragmentation. As regulatory pressures intensify globally, more localized governance structures may enhance platforms' ability to comply with diverse legal regimes, reduce regulatory risk, and maintain operational continuity. In this sense, redistributing decision-making authority is not only normatively desirable but increasingly aligned with platforms' long-term strategic interests.

Developing contextually responsive governance frameworks

Rather than pursuing universal policies that inevitably reflect particular cultural assumptions, platforms should develop governance frameworks that maintain core principles while creating systematic mechanisms for contextual interpretation. A contextually responsive framework requires moving beyond ad hoc, case-by-case judgments by individual moderators toward systematic processes for local adaptation. This includes the development of cultural competency protocols that guide how content is interpreted across different linguistic, social, and political contexts, particularly in borderline cases where meaning is highly contingent.

At the same time, fully rewriting global content policies is neither realistic nor likely. Platforms face strong incentives to maintain standardized rule systems, and recent regulatory developments, such as those in the European Union and the United Kingdom, have shown that while external pressure can reshape enforcement processes and remedies, it has limited impact on the underlying substantive rules.

In practice, a contextually adaptive governance model would consist of three core elements. First, platforms should introduce jurisdiction-specific overlays, allowing for targeted adjustments to rules and enforcement in regions where substantial divergence is required, developed through engagement with both governmental and civil society actors. Second, these frameworks should be implemented in tandem with greater decision-making authority for regional teams, ensuring that contextual interpretation is embedded in enforcement practices. Third, platforms should invest in improving the training and calibration of automated systems, including the development of localized datasets and evaluation benchmarks, to enhance the accuracy and consistency of moderation across diverse contexts.

Across these elements, the content–conduct–context matrix can serve as a guiding tool, helping platforms systematically assess how different dimensions of ambiguity manifest across jurisdictions and how policy responses should be adapted accordingly.

Transforming labor conditions and professional pathways

The exploitative vendor relationships that dominate content moderation undermine both worker welfare and governance quality, creating systems where those doing the most culturally sensitive work have the least authority and professional development opportunities. Addressing this imbalance requires rethinking how moderation labor is organized, valued, and integrated into governance processes. First, platforms could move beyond purely extractive vendor relationships by creating clear career pathways for moderators, including progression into policy, quality assurance, and trust and safety strategy roles. Cultural and contextual expertise should be recognized as a specialized professional competency, rather than treated as interchangeable labor.

Second, platforms should establish mechanisms for meaningful participation of moderators in policy development, particularly in areas involving borderline content. This could include structured feedback channels, formal escalation inputs into policy revisions, and the incorporation of moderator insights into auditing and evaluation processes. As frontline interpreters of ambiguity, moderators

are uniquely positioned to identify gaps, inconsistencies, and unintended consequences in existing rules.

Third, improving labor conditions is essential for enhancing governance outcomes. This includes fair compensation, reduced reliance on precarious contracts, better psychological support, and realistic performance metrics that do not penalize careful judgment in complex cases. Without these changes, moderators will continue to default to risk-averse or mechanical decisions that undermine effective borderline content governance.

Establishing accountability and coordination mechanisms

Current accountability frameworks rely heavily on voluntary initiatives and reputational pressure, which have proven insufficient for addressing systemic inequities or coordinating responses to cross-platform borderline content. More robust mechanisms must combine external oversight with binding enforcement powers and inter-platform coordination capabilities. Building on limited successes of initiatives like the GNI, new frameworks should include mandatory human rights impact assessments for content governance decisions, particularly those affecting marginalized communities or political speech. These assessments should be conducted by independent bodies with authority to require policy changes and remediation when violations are identified. Additionally, the emergence of cross-platform content strategies requires shared intelligence networks that enable coordination across the currently skewed trust and safety market.

Summary

These policy recommendations collectively point toward the need for a fundamental restructuring of content governance, one that centers equity, embraces interpretive complexity, and builds institutional mechanisms capable of serving diverse global communities. Implementing such changes will require sustained commitment, significant resource investment, and a willingness to redistribute power in ways that challenge existing industry structures. However, the alternative — maintaining systems that marginalize the majority of the world's users while concentrating decision-making authority in Western corporate headquarters — is increasingly unsustainable in a digital environment that shapes global communication and democratic participation.

At the same time, the growing politicization of platform governance has made companies more risk-averse, often discouraging the kind of nuanced, context-

sensitive deliberation that borderline cases demand. In response, platforms have tended to default to standardized and defensible approaches that privilege dominant cultural perspectives over locally grounded interpretation.

The three-dimensional framework highlighted in this research offers a pathway toward more systematic and equitable handling of borderline cases. However, realizing this potential ultimately depends not only on technical or procedural reforms, but on platforms' willingness to exercise political will: to redistribute authority, invest in context-sensitive governance mechanisms, and meaningfully center the cultural expertise that effective moderation requires.

Appendix:
Methodological notes

This research combines policy analysis and qualitative interviews to examine the complex ecosystem of borderline content moderation. The methodological design acknowledges that content moderation operates as both a technical and deeply human practice, requiring approaches that can capture both institutional dynamics and lived experiences of practitioners.

Policy analysis:

We analyzed publicly available content policies and guidelines from major platforms including Meta, YouTube, TikTok, Twitter/X, and LinkedIn, focusing specifically on how these companies conceptualize and operationalize borderline content categories. This analysis traced the evolution of borderline content frameworks from their emergence in 2018-2019 through current implementations (August 2025), examining both explicit policy statements and practical applications revealed through transparency reports and public communications.

Expert interviews:

The research draws on in-depth, semi-structured interviews with content moderation practitioners, including human moderators, policy specialists, trust and safety experts, and researchers across the global content governance ecosystem *(see List of Interviewees)*. The sample was purposefully constructed to capture perspectives across different roles and levels of the moderation pipeline, with particular attention to including voices from Majority World contexts that are often underrepresented in existing research. Interview data were analyzed using thematic analysis, focusing on recurring patterns related to power dynamics, institutional structures, and the interaction between local expertise and global policy-making. All participants provided informed consent for participation, recording, and the use of anonymized quotations.

Limitations

This study has several limitations. First, platform policies are dynamic and frequently updated. As a result, the analysis captures a snapshot of governance frameworks at a particular moment in time, and may not fully reflect subsequent policy changes or evolving enforcement practices. Second, while the interview sample includes perspectives from multiple regions, it does not comprehensively cover all geographic contexts. Certain regions and language communities remain underrepresented, which may limit the extent to which the findings capture the full diversity of global moderation practices. Third, content moderation systems vary significantly across platforms depending on their scale, organizational structure, and user base. Smaller or regionally focused platforms may operate under different constraints and priorities than large global platforms. As such, the governance patterns identified in this research may not be fully generalizable across all platform types. Finally, the study focuses on practitioner perspectives and platform policies, and does not include systematic analysis of user experiences or quantitative enforcement data, which could provide additional insights into how borderline content moderation operates in practice.

Despite these limitations, the research offers a grounded and empirically informed account of borderline content moderation, highlighting structural patterns and governance challenges that are consistently observed across platforms and contexts. The findings are intended to provide an analytical framework and set of policy-relevant insights, rather than universally generalizable conclusions.

List of Interviewees

No.	Name	Affiliation	Sector	Role	Expertise	Year of experience in the field
ID1	Hassan A. Naizi	Common Law Chambers: Attorneys & Consultants, formerly Meta South and Central Asia	Industry	Lawyer	Content regulation, freedom of expression, data protection, regulatory compliance	12
ID2	Chanell Daniels	-	Industry	Manager	AI risk management, trust and safety	6
ID3	-	-	Industry	Senior Director - T&S Product	Content policy, trust and safety, operations	20
ID4	Mona Elswah	University of Exeter	Academia, civil society	Researcher	Content moderation	7
ID5	Andy Briercliffe	Independant	Civil society	Online harms consultant / advisor	Online harms, child safety	15
ID6	-	-	Industry, civil society	Trust & Safety investigator	Behavioral and content integrity	10
ID7	-	-	Business Process Outsourcing Industry	Senior executive	Generative AI content moderation	3
ID8	-	-	Business Process Outsourcing Industry	Content moderator	Generative AI & content moderation	3

No.	Name	Affiliation	Sector	Role	Expertise	Year of experience in the field
ID9	Irene Mwendwa	-	Civil society	Tech Policy Lawyer	Technology and data policy	12+
ID10	Disha Verma	Tech Global Institute	Civil Society	Program Manager, Asia Pacific	Technology and rights	5

**Notes: the interviewees would like to clarify that these are their own opinions and do not reflect the institution's position.*

Authors

Diyi Liu is a 2025 Tech Policy Fellow at TGI and holds a DPhil in Information, Communication, and the Social Sciences at the Oxford Internet Institute, University of Oxford.

Shahzeb Mahmood is the Head of Research at Tech Global Institute, specialising in technology law, digital policy, and internet governance. He holds a Master of Laws (LL.M.) from the University of New South Wales and a Master of Public Policy and Management from the University of Melbourne. He is also a barrister of the Honourable Society of Lincoln's Inn.

www.techglobalinstitute.com