13

Content Moderation in Practice

Laura Edelson

13.1 INTRODUCTION

Almost all platforms for user-generated content have written policies around what content they are and are not willing to host, even if these policies are not always public. Even platforms explicitly designed to host adult content, such as OnlyFans,[1] have community guidelines. Of course, different platforms' content policies can differ widely in multiple regards. Platforms differ on everything from what content they do and do not allow, to how vigorously they enforce their rules, to the mechanisms for enforcement itself. Nevertheless, nearly all platforms have two sets of content criteria: one set of rules setting a minimum floor for what content the platform is willing to host at all, and a more rigorous set of rules defining standards for advertising content. Many social-media platforms also have additional criteria for what content they will actively recommend to users that differ from their more general standards of what content they are willing to host at all.

These differences, which exist in both policy and enforcement, create vastly different user experiences of content moderation in practice. This chapter will review the content-moderation policies and enforcement practices of Meta's Facebook platform, YouTube (owned by Google), TikTok, Reddit, and Zoom, focusing on four key areas of platforms' content-moderation policies and practices: the content policies as they are written, the context in which platforms say those rules will be enforced, the mechanisms they use for enforcement, and how platforms communicate enforcement decisions to users in different scenarios.

Platforms usually outline their content-moderation policies in their community guidelines or standards. These guideline documents are broad and usually have rules about what kinds of actions users can take on their platform and what content can be posted. These guideline documents often also describe the context in which

[1] *Help*, Onlyfans, https://perma.cc/WCW7-VDSY.

https://doi.org/10.1017/9781009174411.017 Published online by Cambridge University Press

rules will be enforced. Many platforms also provide information about the enforcement actions they may take against content that violates the rules. However, details about the consequences for users who post such content are typically sparse.

More detail is typically available about different platforms' mechanisms for enforcement. Platforms can enforce policies manually by having human reviewers check content for compliance directly, or they can employ automated methods to identify violating content. In practice, many platforms employ a hybrid approach, employing automated means to identify content that may need additional human review. Whether they employ a primarily manual or primarily automated approach, platforms have an additional choice to make regarding what will trigger enforcement of their rules. Platforms can enforce their content-moderation policies either proactively by looking for content that violates policies or reactively by responding to user complaints about violating content.

Platforms also have a range of actions they can take regarding content found to be policy violating. The bluntest tool they can employ is simply to take the content down. A subtler option involves changing how the content is displayed by showing the content with a disclaimer or by requiring a user to make an additional click to see the content. Platforms can also restrict who can see the content, limiting it to users over an age minimum or in a particular geographic region. Lastly, platforms can make content ineligible for recommendation, an administrative decision that might be entirely hidden from users.

Once a moderation decision is made, either by an automated system or by a human reviewer, platforms have choices about how (and whether) to inform the content creator about the decision. Sometimes platforms withhold notice in order to avoid negative reactions from users, though certain enforcement actions are hard or impossible to hide. In other instances, platforms may wish to keep users informed about actions they take either to create a sense of transparency or to nudge the user not to post violating content in the future.

13.2 FACEBOOK

Facebook (owned by Meta) has made more information about its content-moderation policies and practices available compared to other social-media companies discussed here. However, it is also the only major platform at the time of this writing that gives an outside body, its external Oversight Board, discretion over the enforcement of its policies.

13.2.1 *Content Policies*

Facebook outlines its content policies in its Community Standards.[2] Broadly speaking, Facebook prohibits or otherwise restricts content that promotes violent or

[2] *Facebook Community Standards*, META, https://perma.cc/G36P-CAU8.

https://doi.org/10.1017/9781009174411.017 Published online by Cambridge University Press

criminal behavior, poses a safety risk, or is "objectionable content," usually defined as hate speech, sexual content, or graphic violence.

Violent, sexual, hateful, and fraudulent content are all prohibited outright. However, there are limited exceptions for newsworthy content, such as police body-cam footage from shooting incidents, which must be shared behind a warning label if at all. Content that poses an immediate safety risk, such as non-consensual "outing" of LGBTQ+ individuals or doxing, is always prohibited. Many other forms of "borderline" content are restricted, rather than banned outright, if it is found to be satirical, expressed as an opinion, or newsworthy.

Meta's policy around misinformation is more ambiguous than these prohibited categories of content. The company's policy says, "misinformation is different from other types of speech addressed in our Community Standards because there is no way to articulate a comprehensive list of what is prohibited." The policy continues, "We remove misinformation where it is likely to directly contribute to the risk of imminent physical harm. We also remove content that is likely to directly contribute to interference with the functioning of political processes and certain highly deceptive manipulated media."[3] In practice, this policy has produced subcategories of misinformation with varying levels of protection. For example, over the past several years, the company has interpreted this policy as prohibiting vaccine misinformation but not climate change-related misinformation.

13.2.2 *Enforcement Practices*

Meta also provides some information about Facebook's policy-enforcement practices in its "Transparency Center."[4] Facebook says that it enforces its policies with a mix of automated methods and human reviewers who train the automated systems over time. In Meta's words, a new automated system "might have low confidence about whether a piece of content violates our policies. Review teams can then make the final call, and our technology can learn from each human decision. Over time – after learning from thousands of human decisions – the technology becomes more accurate."[5]

This quote describes a fairly standard process in machine learning where automated systems and humans collaborate to make decisions, with humans having a more significant role early in the process and automated systems "learning" from the decisions humans make over time. While Meta's documentation clearly states that human reviewers make the call when automated classifiers have low confidence, it is less clear about human reviewers' role in more established domains. Meta states that there are some circumstances where automated systems remove content

[3] *Misinformation*, Meta, https://perma.cc/2DTC-R7CT.
[4] *How Meta Enforces Its Policies*, Meta, https://perma.cc/82GV-37N6.
[5] *Id.*

https://doi.org/10.1017/9781009174411.017 Published online by Cambridge University Press

without human intervention: "Our technology will take action on a new piece of content if it matches or comes very close to another piece of violating content." According to Meta, their "technology [i.e., automated system] finds more than 90% of the content we remove before anyone reports it for most violation categories."[6] A careful reader will note that this does not say that 90 percent of content is *removed* before users report it, only that it is *found* before users report it. Still, it is likely a safe assumption that the vast majority of content moderation that happens on the Facebook platform is proactive, rather than reactive.

When Facebook removes content (as opposed to restricting who can see their content or reducing how often it recommends it in users' newsfeeds), it notifies the user who posted the content.[7] It then employs a "strike" system to restrict the accounts of users whom the company finds to have violated content policies repeatedly over time.[8] A first strike is only a warning, but after that, strikes result in increasingly longer bans from creating content. These range from a second strike resulting in a one-day ban to a fifth strike resulting in a thirty-day ban. Users can appeal decisions they think are incorrect, and Meta publishes statistics about how often they reinstate removed content in various categories of violations in its quarterly Community Standards Enforcement Report.[9] Finally, accounts that repeatedly post policy-violating content and thus receive five or more strikes can be disabled entirely.[10] As a final layer of oversight of their content-moderation practices, Meta, uniquely among major social-media companies, has established an Oversight Board.[11] The Board serves, among other things, as a final court of appeals for Facebook's moderation decisions. As of the time of this writing, Meta's Oversight Board has reviewed thirty-six appeals, and found in twenty-four cases that content should be reinstated.[12]

13.3 YOUTUBE

Rather than a standalone section of its website, YouTube outlines its content policies ("Community Guidelines") in a section of its Help pages.[13] YouTube prohibits nearly all the same categories of content as Facebook, although the companies' policies use different nomenclature in some cases and demonstrate different areas of focus. For example, both platforms prohibit sexual content, but Facebook groups this category under the umbrella of "offensive content" while

[6] *How Technology Detects Violations*, META (Jan. 19, 2022), https://perma.cc/QC6Q-L9RM.
[7] *Taking Down Violating Content*, META (Sept. 9, 2022), https://perma.cc/B3VX-388A.
[8] *Restricting Accounts*, META (Oct. 4, 2022), https://perma.cc/A7BJ-AHPF.
[9] *Community Standards Enforcement Report*, META, https://perma.cc/9BHW-SAPP.
[10] *Disabling Accounts*, META (Jan. 19, 2022), https://perma.cc/RYR7-RZ6J.
[11] OVERSIGHT BOARD, https://perma.cc/M32S-356A.
[12] *Id.*
[13] *YouTube's Community Guidelines*, YOUTUBE, https://perma.cc/85SE-MW4X.

YouTube groups it with "sensitive content." Similarly, both platforms broadly prohibit fraudulent content, but YouTube focuses more on preventing spam, while Facebook focuses on financial scams.

In contrast to its relatively well-developed documentation around its content policies, YouTube's documentation[14] of its policy-enforcement mechanisms is sparse. The company thoroughly describes how users can flag content that violates policy and how content is reactively reviewed when that happens (always by human reviewers). The policies state that YouTube does, however, "use technology to identify and remove spam automatically, as well as re-uploads of content we have already reviewed and determined violates our policies."[15] Google (YouTube's owner) also publishes data about content moderation on YouTube in quarterly Transparency Reports.[16] In these reports, Google breaks down the share of removals originating from automated systems versus users, with greater than 90 percent of removals originating from automated systems. Google also provides statistics on when in a post's lifecycle removals happen, breaking down the share that happens before a post receives any views at all, one to ten views, or greater than ten views.

Like Facebook, YouTube employs a "strike" system to nudge users into better behavior.[17] YouTube's strike system is significantly more aggressive, however. Users get a warning with no other penalty attached the first time YouTube finds that they have posted content that violates its policies. After that, users who receive three additional strikes in a ninety-day period will have their YouTube channel permanently removed. YouTube further says that "[i]f your channel or account is terminated, you may be unable to use, own, or create any other YouTube channels/accounts."[18] This implies that channel removal is indeed a complete ban of the user in some cases, but it's unclear how often this penalty is imposed in full.

13.4 TIKTOK

TikTok, similar to Facebook, maintains a separate "Community Guidelines" section of its website.[19] Content prohibitions are grouped slightly differently, but they generally resemble those of other platforms insofar as they focus on sexually explicit content, fraudulent content, and content deemed to pose a safety risk.

TikTok has released very little information about its mechanisms for enforcement, which violations will result in permanent bans, and how many "strikes" users might receive before getting a permanent ban. In 2021, TikTok published a blog

[14] *YouTube Community Guidelines Enforcement FAQs*, Google, https://perma.cc/X3FD-Q7RM.
[15] *See id.* (answering the question "Is flagged content automatically removed?").
[16] *YouTube Community Guidelines Enforcement*, Google, https://perma.cc/EAS7-X6NQ.
[17] *Community Guidelines Strike Basics on YouTube*, Google, https://perma.cc/6WPD-B2R3.
[18] *Channel or Account Terminations*, Google, https://perma.cc/Y6DC-FZHN.
[19] *Community Guidelines*, TikTok, https://perma.cc/XDM8-DQQ9.

post[20] announcing that the platform would begin automated proactive content removals for some categories of content. The platform also publishes quarterly Community Guidelines Enforcement reports[21] with details around content removal and restoration after appeal.

Unlike Meta and Google, TikTok does not give removal statistics by method of initial flagging. Rather, it breaks down final removals by "automated" versus "manual" means. The word "automated" is undefined, but one can reasonably infer it refers to removals without any human review. In TikTok's case, this appears to be about one-quarter of overall removals, but note that this metric is not equivalent to the ones given by other platforms around initial flagging type, so these numbers are not directly comparable. This is because this metric likely refers to human involvement at any point in the moderation process, instead of solely at the point of initial flagging.

At the same time as its automated proactive-content-removal announcement, TikTok also confirmed that it employs a strike system to ban users who repeatedly post violating content. TikTok does not currently disclose how many times (or at what frequency) users would have to violate policy to receive a ban. Its Community Guidelines make clear that they have a zero-tolerance policy for the most serious categories of violations, such as Child Sexual Abuse Material (CSAM) or violent content. In its transparency reports, the company provides data about the number of accounts removed on a monthly basis. Still, there is no way to connect the number of removed posts to the number of removed accounts without more intermediate data.

13.5 REDDIT

Like other platforms reviewed in this chapter, Reddit publishes Community Guidelines that apply across the entire platform.[22] However, these Community Guidelines are best thought of as a content-moderation "floor" that describes a substantially lower threshold than is actually enforced across the vast majority of the platform. This is because all Reddit content is posted to "subreddits" (also known as channels), each having its own set of policies and practices that users create and enforce themselves.[23] Reddit does require that channel moderators post their policies clearly and maintain an appeals process, but communities are otherwise free to self-moderate as they see fit.

This overarching policy of relatively few limitations on what content is permitted on the platform has naturally led to the existence of many groups with a great deal of

[20] Eric Han, *Advancing Our Approach to User Safety*, TikTok, https://perma.cc/V7Y2-ZG9Y.
[21] *Reports*, TikTok, https://perma.cc/L7YF-4KRF.
[22] *Reddit Content Policy*, Reddit, https://perma.cc/3A9D-3BJ7.
[23] *Moderator Code of Conduct*, Reddit (Sept. 8, 2022), https://perma.cc/GYS2-5UUP.

content that many users would find objectionable for one reason or another. To manage this issue, Reddit has a policy of "quarantining" subreddits that most users might find highly offensive or upsetting.[24] Reddit will not run ads on quarantined channels, which means they generate no revenue for Reddit. Content posted in these channels also does not appear in feeds of users not subscribed to the quarantined subreddits and will not be discoverable in user searches.

Similar to other platforms we have discussed, Reddit publishes a transparency report with details about its content-policy enforcement. However, it only publishes this report annually.[25] Reddit has some site-wide enforcement of its content-moderation policies, but subreddit moderators do the majority of content removal, according to its transparency report. To support the enforcement of both site- and community-specific content guidelines by moderators, Reddit makes an extensive set of moderator documentation[26] and tools[27] available to its army of volunteer channel moderators. One community moderation tool unique to Reddit among the platforms we have discussed is that of *flair*.[28] Flair are short text tags with single words, phrases, or emoticons. While flair can be used for a variety of purposes, when it is associated with user accounts, it typically conveys a user's reputation.

Due to the fragmented nature of both content policy and enforcement on Reddit, there is little that can be said about how enforcement decisions are communicated to users when they happen on the channel level. However, while subreddit moderators have broad autonomy to police their channels (and to ban users from them) as they see fit, only Reddit can ban user accounts from the site entirely. Reddit publishes data about both content and user-account removal in its transparency report, but the platform does not outline any explicit thresholds of policy violations (either what kind or how many) that would prompt a user's account to be suspended.

13.6 ZOOM

While Zoom is not generally considered a social-media company, it is still a platform for users to share content. Readers may be most familiar with Zoom as a tool for one-on-one video calling, but Zoom can also be used to host multi-party calls with up to 1,000 participants and webinars with up to 10,000, depending on the host's account type.[29] Zoom users can also record videos and save them to Zoom's cloud so that others can watch those videos at a later time. Therefore, the company

[24] *Quarantined Subreddits*, REDDIT, https://perma.cc/2FPP-66FQ.
[25] *Transparency Report 2021*, REDDIT, https://perma.cc/7HLX-BT2J.
[26] *Reddit Mods*, REDDIT, https://perma.cc/5HU2-DVRU.
[27] *Reddit Moderation Tools*, REDDIT, https://perma.cc/99P4-T8C3.
[28] *User Flair*, REDDIT, https://perma.cc/49JR-2M7W.
[29] Ajaay, *Zoom Limit: Maximum Participants, Call Duration, and More*, NERDS CHALK (Oct. 21, 2020), https://perma.cc/EWQ8-4YMM.

has published standards for what content it is and is not willing to host.[30] In their community standards, Zoom prohibits many of the same content categories as other platforms we have reviewed. These prohibited categories include hate speech, promotion of violence, and sexual or suggestive content, though some other commonly prohibited categories, such as misinformation, are allowed. However, unlike the other platforms we have discussed, Zoom only enforces its policies in reaction to user reports.[31]

Zoom appears to have no proactive enforcement of its content policies. Zoom also states that all moderation in response to user reports is done manually, rather than by automated means.[32] Notably, the company does not currently publish data about its content-policy enforcement. Instead, Zoom's annual transparency report only includes statistics about the company's responses to government requests of different types. The company has not made data available about how many pieces of content it has removed or how many users have been banned due to its content-policy enforcement.

Zoom does not have external oversight of its content-moderation decisions – only Meta does this – but interestingly, the platform does have several progressive tiers of internal content-moderation review to which users can appeal decisions. At the highest tier of review, an "appeals panel" makes decisions by majority vote. Panel members are chosen from a pool of Zoom employees and serve for no longer than two years. Panel decisions are documented so they can guide future internal decision-making. In many respects, Zoom's "appeals panel" is described quite similarly to Meta's Oversight Board.

13.7 DIFFERENCES IN CONTENT-MODERATION POLICY

Of the platforms we have reviewed, it is likely no coincidence that the three largest – Facebook, YouTube, and TikTok – have similar written policies on content moderation, as they are all attempting to serve very broad user bases and therefore face similar challenges. They all have platform-wide policies against many of the same types of content. They all take tiered approaches to enforcement, involving banning some kinds of content and limiting access or distribution of other kinds of content. They all describe (in greater or lesser detail) a policy of warning users who post violative content and banning those users who do so repeatedly.

Reddit's channel-specific approach is different in almost every respect from the approach taken at Facebook, YouTube, and TikTok. While there is a minimum standard for allowable content on Reddit, most policy rules are set by users themselves to facilitate the types of discussions they want to engage in within specific

[30] *Acceptable Use Guidelines*, Zoom, https://perma.cc/3SS4-86GN.
[31] *Acceptable Use Guidelines Enforcement*, Zoom, https://perma.cc/P8GZ-BKRF.
[32] *Our Tier Review System*, Zoom, https://perma.cc/25TT-JWKD.

groups. As they are written, Zoom's content policies fall somewhere between the permissiveness of Reddit and the broad prohibitions against offensive content that the largest platforms have. Zoom prohibits sexual and fraudulent content, as well as explicit calls for violence. However, the platform makes no explicit rules against many other categories of content, including misinformation, that are harder to define. In this respect, Zoom's content policies are significantly less aggressive than those of Facebook, TikTok, and YouTube.

13.8 DIFFERENCES IN CONTENT-MODERATION ENFORCEMENT RULES

The starkest differences between the platforms we have studied exist not in their policies as they are written, but in their rules for enforcing these policies. For example, Zoom's clear statement that it only enforces its policies in response to user reports creates manifestly different conditions for what content is allowed than exists on platforms that engage in proactive enforcement.

There are also meaningful differences between what consequences platforms impose on users who violate platform rules. Most platforms we have discussed employ "strike" systems of some kind, but not all are clear about what penalties will be enforced after which strike, or how long strikes will be counted. YouTube's clarity on these points is a notable exception. This ambiguity is likely strategic, giving platforms the freedom to adjust their policies in reaction to events without having to communicate every change publicly. It is interesting to note that one of Reddit's rules for its channel moderators is not to create "Secret Guidelines"[33] that aren't clearly communicated to users, even though Reddit itself is largely opaque about how it enforces its own guidelines.

Reddit and Zoom take a much more reactive approach to content moderation than Facebook, YouTube, and TikTok. Reddit, as discussed above, leaves most aspects of content moderation – including enforcement – to its user community. Zoom's content policies look much more like those of Facebook, YouTube, or TikTok on paper, but unlike those platforms, Zoom intervenes only in response to user complaints. In effect, then, any given group of users on a Zoom call can effectively agree on and enforce a local content-moderation policy – much as if they were on a subreddit. Unlike Reddit, however, there is no "floor" of allowable content for consenting users, because Zoom only enforces its content policies if it receives a complaint.

However, there do appear to be some areas where the effects of policy enforcement are relatively consistent across platforms, even if the mechanisms for achieving this effect differ. This is particularly true around content that is simply illegal, such as violent terrorist imagery or CSAM (Child Sexual Abuse Material). Every platform

[33] *Moderator Code of Conduct*, *supra* note 23.

https://doi.org/10.1017/9781009174411.017 Published online by Cambridge University Press

we have discussed here makes clear that not only is this type of content prohibited, but that posting this type of content will result in users losing their accounts immediately, without strikes or warnings.

13.9 DIFFERENCES IN CONTENT-MODERATION ENFORCEMENT IMPLEMENTATION AND TRANSPARENCY

Differences around policy enforcement extend beyond rules for what policy enforcement looks like and what triggers it. There are also serious differences in platforms' implementation of enforcement systems. Zoom's all-manual, tiered enforcement system has very different accuracy characteristics than systems that use machine learning to evaluate content proactively. TikTok appears to rely more heavily on fully automated content moderation with an expectation that users will dispute some decisions and some content will be restored after those disputes. These details of implementation create very different user experiences than exist on other platforms.

Some of these differences are the result of platforms' differing structures. Reddit's uniquely manually intensive moderation system results from its channel-focused design. Reviewing the resources needed to build accurate machine-learning systems is beyond the scope of this chapter. However, the largest platforms that employ machine-learning techniques to identify violative content in an automated manner can do so, at least in part, because of the enormous training sets of data they can build because of the large volumes of user content they host.

All of the platforms we have reviewed publish transparency-report documents that provide some information about how their policies are implemented in practice. Each of these "transparency reports" have developed independently and, even when theoretically reporting data about the same category, often use different metrics to measure slightly different things. This means that while they can be individually informative, they are rarely directly comparable.

13.10 CONCLUSION

The platforms reviewed here have profound differences in content-moderation policy, rules for enforcement, and enforcement practices. How, then, can we compare them when they differ on so many dimensions? Ultimately, platforms (and their policies) exist to shape their user experience. This chapter, therefore, proposes that users' ultimate experience of platforms' content policies provides the most meaningful basis for comparison. This outcome-focused framework leads us to a series of questions that can be asked about different categories of content on each platform:

What content are users able to post?
What content will be taken down after users post it and how quickly will it be removed?

https://doi.org/10.1017/9781009174411.017 Published online by Cambridge University Press

What content will be visible to users other than the poster?
What content will be recommended to other users?
What will the consequences be for users who post violating content?

An example of how to apply this framework to a category of content, in this case sexual content, is shown in the table below.

Sexually Explicit Content	Facebook	YouTube	TikTok	Reddit	Zoom
Can users post this content?	May be blocked at time of upload	May be blocked at time of upload	May be blocked at time of upload	Yes	Yes
Will this content be taken down?	Yes	Yes	Yes	Only if it goes against the rules of the channel in which it is posted	Only if a viewer objects
Will this content be visible to other users?	Generally no (because it will not be recommended)	Yes, until it is taken down	Generally no (because it will not be recommended)	Yes, unless it violates channel rules and is removed by a moderator	Yes, unless a viewer objects and the content is taken down
Will this content be recommended to other users?	No	No	No	Only if the user has subscribed to the channel	No. (Zoom does not recommend content)
What are the consequences for users who post this content?	One strike (out of an unknown number)	One strike (out of three to four)	One strike (out of an unknown number)	May be banned from channel (if in violation of channel rules)	Unclear

Platforms and policymakers often discuss aspects of content moderation in isolation. Our exploration of moderation policy and implantation demonstrates the degree to which these dynamic systems are the result of multiple interlocking parts, where aspects of one part of the system impact the efficacy of another. The reality of how policies are experienced by users is heavily impacted by how those policies are implemented. In closing, we encourage the reader, when attempting to make comparisons between platforms or even attempting to understand the impacts of changes to a single system, to consider the whole, rather than the parts.