Securing Federated Platforms: Collective Risks and Responses

Yoel Roth and Samantha Lai

Abstract. As the social media landscape undergoes broad transformation for the first time in over a decade, with alternative platforms like Mastodon, Bluesky, and Threads emerging where X has receded, many users and observers have celebrated the promise of these new services and their visions of alternative governance structures that empower consumers. Drawing on a large-scale textual analysis of platform moderation policies, capabilities, and transparency mechanisms, as well as semi-structured group interviews with developers, administrators, and moderators of federated platforms, we found that federated platforms face considerable obstacles to robust and scalable governance, particularly with regard to persistent threats such as coordinated behavior and spam. Key barriers identified include underdeveloped moderation technologies and a lack of sustainable financial models for trust and safety work. We offer four solutions to the collective safety and security risks identified: (1) institutionalize shared responses to critical harms, (2) build transparent governance into the system, (3) invest in open-source tooling, and (4) enable data sharing across instances.

1 Introduction

Many discussions about social media governance and trust and safety—among regulators, developers, researchers, and users alike—are focused on a small number of centralized, corporate-owned platforms: Meta's Facebook and Instagram, YouTube, X (formerly known as Twitter),[1] Reddit, and a handful of others. The emergence and growth in popularity of federated social media services introduces both new opportunities, and significant new risks and complications. Centralized and decentralized platforms share a common set of threats from malicious users—and require a common set of investments to ensure trustworthy, user-focused outcomes.

While federated services continue to be dwarfed in size in comparison to platforms like Facebook and X, the steady rise in their adoption warrants further attention and study. In the case of Mastodon, for example, changes in ownership and governance at X appear to have significantly accelerated the platform's adoption, with some estimates

[1] Henceforth, the platform will be referred to as "X" in relation to its present-day activity. "Twitter" will be used to refer to the company's actions and decisions prior to November 2022.

10.54501/jots.v2i2.171

showing more than 14 million currently active users (Mastodon, n.d.). Bluesky, launched in February 2023, reached two million users within its first ten months (The Bluesky Team 2023), despite not offering open sign-ups to the public. Threads, Meta's competitor to X, was launched with the promise of integration with decentralized networks (Mohammad, Jarenwattananon, and Detrow 2023; J. Chen 2023), and surpassed 100 million user sign-ups in its first week (Duffy 2023; Roth 2023).

Broadly speaking, federated or decentralized social media refers to a wide array of distinct products, services, and platforms that interconnect using a set of shared communication protocols such as the W3C standard ActivityPub or the under-development Bluesky AT Protocol (ActivityPub, n.d.; Bluesky, n.d.; Graber 2020).[2] In contrast to a centralized social media platform like X, where a user's interactions and content are hosted, managed, and distributed by a single entity that provides both software and business services (like moderation), a federated alternative might involve dozens, hundreds, or even thousands of individual servers running instances (i.e., installations) of an open-source product. Despite being maintained by separate people or groups, servers using the same underlying protocol are interoperable, communicating with each other, and in turn, allowing their users to access each others' content. A number of distinct products have been built atop decentralized protocols, including Mastodon (an X-like social media platform), Pixelfed (an Instagram-like platform focused on media sharing), and PeerTube (a YouTube-like social media platform)—all based on the ActivityPub protocol, and therefore all interoperable with each other. While most federated platforms are noncommercialized and reliant on crowdfunding, the space has also seen newer, commercialized entrants with more financial backing.

These emergent distributed and federated social media platforms offer the promise of alternative governance structures that empower consumers and, optimistically, can help rebuild social media on a foundation of trust. Decentralization enables users to act as hosts or moderators of their own instances (or instances set up for specific or smaller communities), rather than residing within platforms designed with the aspiration of connecting all of humanity—with the goal of increasing user agency, autonomy, and ownership. Governance decisions shift from being made by a single entity responsible for every user on the network to more localized choices made by the administrators and moderators of a user's chosen instance. And, in the event a user objects to the decisions made by one instance in particular, account portability and platform interoperability give users the ability to freely engage with a wide array of product alternatives and instances without having to sacrifice their content or networks.

Alongside these aspirations to empower users and foster more robust, accountable forms of social media governance, federated platforms continue to grapple with many of the same safety and security challenges impacting their centralized counterparts. Nearly two decades of experience with content moderation on social networking platforms like Facebook and X underscores the need for a comprehensive and effective moderation approach for any platform playing host to public conversations. But how does the work of counteracting malign conduct happen in the context of federated social networks?

[2] The terminology used to describe federated and decentralized social media is complex and the subject of some debate. The commonly used term "Fediverse" (as a capitalized, proper noun), for example, refers primarily to the set of platforms built on top of ActivityPub and ActivityStreams, two specific protocols that enable federation. Proponents of these technologies generally argue against using the term "Fediverse" (either capitalized or not) to describe other federated platforms and technologies built on different or competing technical standards, such as Bluesky's AT Protocol and Nostr. This paper's commentary is meant to apply to the broad landscape of federated and decentralized social media platforms, and as such, we use the blanket term "federated and decentralized social media" in place of a more specific term like "Fediverse." While we believe that the underlying analytic conclusions in this paper apply similarly across technologies (both within the ActivityPub Fediverse and beyond it), we recognize that moderation experiences on federated platforms differ depending on scale and audience.

Can existing and nascent structures of content moderation and governance in federated platforms adapt to the specific demands of countering information operations? What structures are necessary to help instance operators, moderators, and the users of federated services address the risks and threats created by persistent adversarial behavior?

This paper offers an empirical assessment of the trust and safety capabilities of federated platforms. Drawing on a broad-based textual analysis of platform communications, blog posts, public code repositories, and social media discussions, we offer a novel comparative assessment of the specific moderation capabilities and technologies (rather than just the stated policies and moderation aspirations) of several of the most prominent centralized and decentralized social media platforms. Rather than evaluating moderation capabilities generally, or focusing on the outcomes of moderation processes (i.e., the presence or absence of harmful content on a given service), we map out the specific structures and capabilities of federated platforms to moderate, and the conditions under which those capabilities may be particularly effective or ineffective.

While we evaluate platform capabilities broadly, our analysis is particularly focused on the policies, technologies, and practices employed by federated platforms to identify and mitigate what we term "collective security risks"—that is, malicious conduct like spam, coordinated manipulation, and inauthentic behavior. We chose to focus on collective security risks for two reasons: first, because of the broad societal and political impact such tactics have had in the past; and second, because these threats require moderation strategies that differ from content-level detections and mitigations discussed in most research about trust and safety. However, despite our initial focus on a subset of social media threats, we find that many of the challenges and gaps identified in this analysis are applicable to other content moderation domains, including efforts to combat child sexual exploitation and the spread of misinformation. Where applicable, we note these similarities in our analysis.

Additionally, we supplemented our textual analysis with a series of semi-structured interviews hosted with a group of 32 platform developers, maintainers, researchers, and moderators, representing experience with or contributions to the development of eight different platforms and federated technologies. These discussions focused on exposing how platform developers, maintainers, and moderators understand the challenges of moderating manipulative behavior and other collective security threats, and what solutions developers and instance moderators themselves believe would be most effective in addressing safety and security threats.

Across most dimensions evaluated, we find that federated platforms have less developed, robust, and scalable capabilities for content moderation than their centralized counterparts. Unique and acute architectural constraints inhibit their ability to defend against common social media behavioral threats. Meanwhile, substantial financial and institutional obstacles exist to developing more scalable content moderation capabilities. We conclude by offering four potential interventions to mitigate collective security risks in the context of federated social media: (1) institutionalize shared responses to critical harms, (2) develop frameworks for transparent and trustworthy governance of collective moderation systems, (3) support open-source tool development, and (4) enable data-sharing across instances.

2 From centralized to decentralized moderation

Content moderation refers to "the organized practice of screening user-generated content… in order to determine the appropriateness of the content for a given site, locality, or jurisdiction" (Roberts 2022). While moderation often has multiple objectives, including compliance with various laws, its foundational purpose is typically to protect the users of a platform—and, potentially, other people and groups impacted by on-platform behavior—from a range of online harms (Wardle and Derahkshan 2017; Gillespie 2018; Thomas et al. 2021). While these now-ubiquitous practices are increasingly professionalized and institutionalized (Slater and Masiello 2023), forms of content governance have existed in various forms for decades, carried out both by individuals and leaders within online communities and by platform staff (Matias 2016). As platforms have grown and commercialized, moderation has become entrenched as an essential business practice: In the context of commercial platforms, their "economic viability depends on meeting users' speech and community norms...if a platform creates a site that matches users' expectations, users will spend more time on the site and advertising revenue will increase" (Klonick 2017).

Given the wide range of threats, tactics, and technologies that fall under the umbrella of social media content moderation, we find it helpful to break down moderation into three discrete components, drawing on Camille François's "ABC" taxonomy: Actors, Behaviors, and Content (François 2019).

Nearly all content moderation discussions begin with an assessment of the *content* of a post or account: the language it uses, the links it shares, and the information a user chooses to share on their profile. These assessments treat content moderation as a challenge of evaluating and making decisions about what a post or account says or appears to be. Typically, content-focused moderation tasks include the removal of illegal content that platform users would not want to see, such as child sexual abuse material (CSAM) and terrorist content, as well as the management of other forms of potentially or perceptually harmful or unwanted content such as nudity, pornography, or material that promotes harmful behaviors such as self-harm, eating disorders, or suicide (Spence et al. 2023; Roberts, Wood, and Eadon 2023; Levine 2022; Suzor, Seignior, and Singleton 2017; Gorwa, Binns, and Katzenbach 2020; Chancellor et al. 2016).

However, content-level analyses only reveal part of the picture of content moderation's necessary scope and scale. Platforms also commonly moderate manipulative or disruptive *behaviors* (as François (2019) terms them), including the creation of inauthentic accounts, bulk and high-volume posting, and manipulation of engagement metrics (such as likes or reposts). Common outcomes of manipulative behaviors include spam, financial scams, coordinated inauthentic behavior, and phishing (Lukito 2019; Arif, Stewart, and Starbird 2018; Ong and Cabañes 2018; Koutrika et al. 2008). These practices, while commonplace and often innocuous (as in the case of commercially motivated spam), can substantially alter the stakes for platform content moderation. Researchers studying platform manipulation campaigns, including those carried out by nation-state actors, have observed that "the raw volume of activity generated by an operation can far outweigh the baseline authentic civic activity in an ecosystem," drowning out oppositional voices and organic user conversations (Conlon, Nuland, and Karan 2022). Especially from 2015 onward, high-volume, low-sophistication political manipulation campaigns became a feature of Twitter in particular (DFRLab 2019; Nimmo, Eib, and Tamora 2019). As François (2019) puts it in the ABC framework: "At the end of the day, deceptive behaviors have a clear goal: to enable a small number of actors to have perceived impact that a greater number of actors would have if the campaign were organic."

Evaluating behavior as a component of moderation, and integrating behavioral signals into platform moderation strategies, is essential for accounting for some of the most pernicious and impactful threats to online discourse. Many of the malign campaigns targeting social media are difficult, if not impossible, to identify based on content alone. Looking back at key examples of the Russian Internet Research Agency's activity on

Twitter in 2016, what's most striking about posts from prominent accounts like "Crystal Johnson," a Russian persona purporting to be an African-American woman, is that, by and large, the content of the posts is true: While the IRA's earliest efforts involved comically ineffective rumormongering about an alleged outbreak of Ebola in Atlanta, the bulk of their activity during and after the 2016 US elections used a more subtle tactic of sharing factually accurate but divisive rhetoric using inauthentic behaviors (such as fake accounts) (Giridharadas 2022; A. Chen 2015). This stymies efforts to moderate social media activity based on policies that evaluate only the content of a post.

Finally, most content moderation approaches involve at least some consideration of the *actors*—that is, people, groups, governments, or other entities—responsible for online activity. Often, actor-level analysis is reduced to the security practice of "attribution"—the name-and-shame exposure of the individuals or groups responsible for an attack or intrusion (Rid and Buchanan 2014)—but attribution is far from the only goal of actor-level analysis. Understanding the actors responsible for malicious activity meaningfully influences how platforms respond to these threats—and can give platforms necessary tools for addressing these challenges in a scalable way.

In particular, a key practice employed by a number of platforms as part of their moderation efforts is longitudinal analysis of specific threat actors—that is, tracking and analyzing the behavior of specific individuals or groups, or the patterns of malicious behavior over time. These practices have typically been carried out both by platforms themselves, and by a wide array of civil society and academic groups (Nimmo and Agranovich 2022; Bradshaw et al. 2021; Graphika Team 2020; Graphika and Observatory 2023; Linvill and Warren 2020; Arif, Stewart, and Starbird 2018). Understanding the behaviors and motivations of persistent threats helps platforms develop effective mitigation strategies suited to applying optimal, cost-effective pressure to a particular actor based on their unique goals and constraints.

Alongside moderation strategies integrating assessments of actors, behaviors, and content on social media sites, *transparency* initiatives have emerged as core components of platform moderation efforts. Various transparency mechanisms were first voluntarily adopted by many centralized platforms beginning in 2010 (Infantino 2013; X 2022; Access Now, n.d.), starting with the publication of periodic quantitative reports about platform moderation decisions. These practices are now increasingly mandated under a number of global regulations (European Commission 2022; Library of Congress 2021; European Union 2022). While the specific contents of transparency reports vary, they are meant to illustrate what actions platforms are taking to mitigate harms on their platforms (Tworek and Wanless 2022). Support for platform transparency is broad-based across stakeholders in government, academia, industry, and civil society, with transparency proponents often arguing that greater visibility into platform actions can contribute to accountability to their users, external scrutiny by independent researchers, and ultimately more informed government policymaking (Keller and Leerssen 2020; MacCarthy 2022; Puddephatt 2021; Gorwa, Binns, and Katzenbach 2020).

As federated and decentralized platforms have grown in usage and popularity, academic scholarship has begun to examine the governance and trust and safety dimensions of their development and adoption. A growing body of literature on decentralized platforms examines their possibilities for introducing new forms of online governance (Masnick 2019c; Mansoux and Abbing 2019; Zuckerman and Rajendra-Nicolucci 2020; Gehl and Zulli 2022; Ermoshina and Musiani 2022). Masnick (2019c), for example, notes that federation can "push the power and decision making out to the ends of the network, rather than keeping it centralized among a small group of very powerful companies."

A key finding in early studies of federated and decentralized platforms has been that, despite a lack of protocol-mandated governance, many federated platforms nevertheless engage in at least some form of moderation—but that the manner by which they moderate differs substantially from strategies employed by centralized platforms. The essential feature of federated systems, and of the protocols like ActivityPub underlying them, is decentralization: Each instance of a federated service can choose for itself what its governance approach will be, and, in turn, its governance decisions extend only so far as the (virtual) boundaries of that particular server. As Rozenshtein (2022) puts it, "No instance can control the behavior of any other instance, and there is no central authority that can decide which instances are valid or that can ban a user or a piece of content from the ActivityPub network entirely. As long as someone is willing to host an instance and allow certain content on that instance, it exists on the ActivityPub network." By design, the perimeter of federated platforms is highly permeable; new platforms and users can enter and exit federated systems readily, to both the benefit and detriment of the overall network.

Rozenshtein (2022) emphasizes that even as some of the tools and technologies used by federated platforms to moderate user behavior overlap with the capabilities of centralized platforms, federated moderation fundamentally differs from centralized moderation by virtue of the distribution of accounts across multiple instances. A user account has a "local" instance on which it resides, and that instance's moderators have the ability to take direct, destructive action on the user's content (such as deleting it). But, for nonlocal accounts and content—that is to say, users whose accounts reside on other instances—the administrators of interoperating instances can only impact their local copies of that content, which influences only the experiences of their local users. If a user on instance A encounters a harmful post from a user on instance B, instance A's moderators have little ability to compel instance B to take any action on the harmful post.

This structural reframing of moderation as a local (i.e., instance-level) rather than network-level decision has the beneficial effect of giving users greater choice about the policies and governance approaches influencing what they see on social media, but it also makes it more challenging to address risks on federated platforms created by instances that either cannot or choose not to moderate. The harmful effects of nonlocal content may persist through on- and offline action by instance users even if they are cordoned off from other parts of federated platforms through localized blocks.

Existing studies of federated governance have noted that these new structures introduce their own challenges, including high costs associated with moderation and the complications to collective governance inherent to a decentralized network, where moderators would face limited means in holding badly moderated or malicious instances accountable (Keller 2021; Rozenshtein 2022; Struett et al. 2023). It is on these challenges in particular that we focus our analysis.

3 Methods

As part of this work, we conducted (1) a census of available texts to map out the existing capabilities of platforms; and (2) semi-structured group interviews with moderators, administrators, practitioners, and researchers working on and with federated platforms.

Drawing on approaches employed by Cramer et al. (2023) and Nicholson, Keegan, and Fiesler (2023) to evaluate the trust and safety engineering and community standards development processes of federated platforms, we conducted a large-scale textual analysis to inform comparisons of trust and safety capabilities across centralized and decentralized platforms. We defined a taxonomy of platform moderation capabilities and the effectiveness of capability implementation, ranging from “None” (the capability has not been implemented) to “Complete” (based on available documentation, the capability exists and operates at the platform’s requisite scale). The full taxonomy is listed in Appendix C. Through internet searches and reviews of platform documentation (such as help centers and support websites), we identified texts of interest, then archived and stored them to capture a point-in-time assessment of platform capabilities. Three independent reviewers evaluated the collected texts against the predefined effectiveness taxonomy to code platform capabilities. In cases where there was inter-coder disagreement, we made efforts to contact platform developers or practitioners for additional information to arrive at a clearer conclusion; where information was not available or conclusive, we erred toward labeling as partial with footnote clarifications.

We also conducted three semi-structured group interviews to solicit qualitative perspectives about threats, platform moderation capabilities, and community-identified mitigation measures. Across the three interviews, there were 32 participants, representing 13 developers, administrators, and moderators of federated or decentralized platforms, 13 researchers affiliated with civil society organizations, and six academic researchers. In their research, development, or moderation practices, interviewees represented a total of eight distinct platforms, including Mastodon, Hometown, Bluesky, Nostr, Mozilla.social, Twitter, Dreamwidth, and Facebook.

Interviews and discussions were focused around three broad themes: (1) community risks and responses, (2) technological risks and responses, and (3) institutional needs and responses. Specifically, we prompted interviewees with specific questions (see Appendix A for a list of interview questions), and encouraged interviewees to engage with us and with each other in an open-ended way in line with the interview’s theme. Following the interviews, transcripts and readouts from the interviews were provided to participants, who had the opportunity to provide additional commentary and feedback. We performed a textual analysis of transcripts from both small group and large group interviews, and incorporated them into the study’s findings.

4 Assessing federated moderation capabilities

We performed a comparative evaluation of the moderation capabilities of centralized and federated platforms across three broad areas: (1) policy, (2) enforcement technologies and capabilities, and (3) transparency.

While a comprehensive assessment of the policies of federated platforms (including the legitimacy of those policies, and their sufficiency in protecting speech and user safety) is beyond the scope of this article, we found that where platform-specific policies do exist, the community standards of federated platforms’ instances are often sparse, high-level statements of principle, rather than the detailed policies published by larger, centralized platforms (see Table 1 on the following page). This creates practical ambiguities for the people responsible for moderating content, as well as uncertainty for users about precisely what is permissible in a given context.

To implement these policies, most federated platforms provide instance administrators and moderators with a rudimentary set of moderation tools and capabilities (summarized in Table 2 on page 9). These capabilities most commonly include the ability to delete individual pieces of content and restrict or ban accounts—though in some cases, platform

Table 1: Platform policies state assessment (see Appendix C for sources)

	Centralized						Federated (future)[a]		Federated (current)			
Capability	Facebook	Instagram	Horizon Worlds	X	Reddit	YouTube	Threads	Bluesky	Mastodon	Pixelfed	diaspora	PeerTube
Publicly available community standards	✔	✔	✔	✔	✔	✔	✔	✔	○	✔	✔	✔
Publicly available specific policy definitions and enforcement criteria	✔	✔	×	✔	×	✔	○	○	×	○	×	○
Platform manipulation/behavioral policies and enforcement criteria	✔	✔	×	✔	○	✔	○	×	×	×	×	×

× None: The capability does not exist.

○ Partial/Likely: The capability exists, but (1) is not applicable to all of the platform's core products/business units, or (2) has significant functionality gaps that prevent effective use for moderation. Alternatively, the platform likely possesses the capability, but does not have publicly listed information on it.

✔ Complete: The existence of the capability is publicly documented, and is available for use (or has demonstrably/documentably been used) at scale across the platform's core products/business units.

N/A Not applicable based on the features/design of the platform.

[a] This refers to platforms that are currently centralized but have promised future integration with other federated services. The following assessment only considers their centralized trust and safety capabilities. Whether these capabilities translate into a decentralized context remains unknown.

developers have implemented more advanced capabilities like media hashing.[3]

Virtually all platforms, including each of the federated platforms we examined, give users some ability to report content they believe may be harmful or in violation of the platform's policies. These reports are submitted to instance administrators and moderators, who then review and (potentially) enforce upon them. However, while reporting features exist, these capabilities lack some of the functions critical for fostering community engagement with moderation (Vilk and Lo 2023). Many lack polymorphic (i.e., content-agnostic) reporting options, which would allow users to report URLs, media, and hashtags that might be problematic or harmful.

In addition, federation introduces unique complications for the processing and management of reports. Reports are largely instance-specific, and given the localized nature of federated moderation, generally do not include a path to making a broader network of instance administrators aware of shared threats. Instance moderators have developed their own methods for informal communication across instances, though it remains challenging for them to engage with each other in a structured way to counteract shared threats. For example, on Mastodon, tagging discussions using "#fediblock" has emerged as a grassroots practice for sharing information about bad actors (thefuturebird 2023), but these approaches have run up against the challenges of a fully distributed, low-trust model: moderators report that it's hard to know which accounts have engaged in sufficiently bad behavior to warrant enforcement without firsthand confirmation.

These same challenges extend to defederation, a novel enforcement capability federated platforms have developed specifically to address the risks of harmful conduct by nonlocal, federated instances. Most federated services offer administrators the ability to take moderation action at the instance level, impacting all users on a remote instance, instead of just moderating post by post or account by account. In the case of Mastodon, for example, instances are able to defederate themselves from other servers—in essence, refusing to communicate with or to display content from a server deemed to

[3] In at least one noteworthy case with Meta's Threads product, centralized and decentralized moderation have commingled as a product of Threads' hybrid approach to federation—raising significant questions about how moderation gets applied across the user experience. In general, moderation tools for on-platform content do not apply directly to content coming in from other federated instances. For example, Threads' existing moderation capabilities only apply to content that is hosted on the platform (i.e., created by Threads users within the Threads app). As the platform opens itself up to interactions with other federated services, the company may need to adapt and build on existing tooling to detect and address threats from other instances.

Table 2: Platform enforcement capabilities state assessment (see Appendix C for sources)

	Centralized						Federated (future)[a]		Federated (current)			
Capability	Facebook	Instagram	Horizon Worlds	X	Reddit	YouTube	Threads	Bluesky	Mastodon	Pixelfed	diaspora	PeerTube
User reporting capabilities for policy violations	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Permanent account bans	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Temporary account bans/timeouts	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	×	✔
Ban evasion detection	○	○	○	○	○	○	○	×	○	○	×	○
Post/content deletion	✔	✔	N/A	✔	✔	✔	✔	✔	✔	✔	✔	✔
Account visibility restriction	✔	✔	N/A	✔	✔	✔	○	✔	✔	✔	×	✔
Post/content visibility restriction	✔	✔	N/A	✔	✔	✔	✔	✔	✔	✔	×	✔
Demonetization	✔	✔	✔	✔	N/A	✔	N/A	N/A	N/A	N/A	N/A	N/A
Automated enforcement tools (heuristics, ML)	✔	✔	×	✔	✔	✔	✔	✔	×	○	○	×
URL blocking	✔	✔	N/A	✔	✔	✔	✔	×	×	×	×	×
Media hashing and matching	✔	✔	N/A	○	✔	✔	✔	✔	×	×	×	×
User-facing moderation controls (block, mute, etc.)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	○	○
User identity verification (ID checks, etc.)	✔	✔	✔	✔	○	✔	✔	×	×	×	×	×
Antispam challenges (reCAPTCHA, phone verification)	✔	✔	N/A	✔	○	✔	○	×	○	○	○	○
Defederation/instance blocking	N/A	N/A	N/A	N/A	N/A	N/A	N/A	×	✔	✔	×	✔

× None: The capability does not exist.

○ Partial/Likely: The capability exists, but (1) is not applicable to all of the platform's core products/business units, or (2) has significant functionality gaps that prevent effective use for moderation. Alternatively, the platform likely possesses the capability, but does not have publicly listed information on it.

✔ Complete: The existence of the capability is publicly documented, and is available for use (or has demonstrably/documentably been used) at scale across the platform's core products/business units.

N/A Not applicable based on the features/design of the platform.

[a] This refers to platforms that are currently centralized but have promised future integration with other federated services. The following assessment only considers their centralized trust and safety capabilities. Whether these capabilities translate into a decentralized context remains unknown.

be problematic, rendering all of its users' content invisible for all the users on an instance that it has chosen to defederate from. Defederation is largely a server-by-server decision, and while there are composite directories of denylists (see, for example, "The Bad Space" (n.d.) and FediSeer (n.d.)), these services presently require extensive involvement from individual moderators, with limited technical or community support. Despite their resource-intensive nature, tactics such as the sharing of denylists have at times been deployed as a form of broad-based, collective action by federated platforms' instance moderators—including the notable case of broad defederation from extremist platform Gab (Masnick 2019a).

While defederation offers one scaled mechanism for addressing repeated or prolific harmful conduct, letting moderators address the behavior of all the accounts residing on a given instance collectively rather than individually, we find that federated platforms largely lack other capabilities for scaled or automated content moderation commonly implemented by centralized platforms. For example, the moderation tools built into platforms like Mastodon do not offer appropriate targeting mechanisms or remediations to moderators that could help them keep pace with high-volume or persistent activity. Moderation actions are wholly manual, and are limited to either banning or restricting individual accounts, or blocking entire ranges of IP addresses or email domains. Both Mastodon and Bluesky, for example, do not provide moderators with the ability to block harmful links from being shared on the service. This prevents moderators from being able to ingest lists of known-bad URLs (such as spam and phishing sites) in order to programmatically restrict them. Mastodon also presently lacks essential tools for addressing media-based harms, such as media hashing and matching functions for addressing child sexual exploitation. Moderators lack the capability to deploy heuristics—essentially, sets of rules that describe patterns of adversarial behavior—that can automate these actions (François 2019; Donovan and Friedberg 2019; Starbird, DiResta, and DeButts 2023; Wanless and Berk 2020), an essential part of the moderation toolkit at all of the existing large, centralized platforms. Although automated content moderation has in certain instances negatively impacted marginalized communities through disproportionate or unjustified takedowns of content, the use of such systems does generally increase the moderators' responsiveness in removing identified categories of problematic content (Griffin 2023).

Finally, most major federated platforms do not have well-established transparency reporting practices—a lack of disclosure that limits the abilities of users and regulators alike to understand the content governance standards and implementations employed by instance operators. Mastodon's 2022 Annual Report, for example, only provides high-level reporting of platform use (Mastodon 2023), rather than the types of granular data about content moderation disclosed by virtually all centralized platforms (see Table 3 on the following page). At present, these shortcomings are largely a matter of user engagement; for all of federation's promise to better empower users to intentionally select instances that align with their values and preferences, users largely lack the information and data they would need to make these decisions in an informed fashion. However, especially as adoption of federated platforms continues to grow, we should expect that regulatory pressures will become an increasingly significant factor. While most federated platforms are still well below the threshold of what the European Digital Services Act would classify as Very Large Online Platforms (VLOPs) subject to transparency reporting obligations, continued growth of these services makes at least some degree of mandatory reporting a seeming inevitability (Komaitis and Franssu 2022).

Table 3: Platform enforcement capabilities state assessment (see Appendix C for sources)

	Centralized						Federated (future)[a]		Federated (current)			
Capability	Facebook	Instagram	Horizon Worlds	X	Reddit	YouTube	Threads	Bluesky	Mastodon	Pixelfed	diaspora	PeerTube
Published transparency report	✔	✔	×	○	✔	✔	×	○	○	×	×	×
Terms of service enforcement data	✔	✔	×	○	✔	✔	×	○	○	×	×	×
Platform manipulation data	✔	✔	×	○	✔	×	×	×	×	×	×	×
Legal information requests data	✔	✔	×	○	✔	✔	×	×	×	×	×	×
Legal removal demands data	✔	✔	×	○	✔	✔	×	×	×	×	×	×
Country or jurisdictional breakdowns of data	✔	✔	×	○	✔	✔	×	×	×	×	×	×

× None: The capability does not exist.

○ Partial/Likely: The capability exists, but (1) is not applicable to all of the platform's core products/business units, or (2) has significant functionality gaps that prevent effective use for moderation. Alternatively, the platform likely possesses the capability, but does not have publicly listed information on it.

✔ Complete: The existence of the capability is publicly documented, and is available for use (or has demonstrably/documentably been used) at scale across the platform's core products/business units.

N/A Not applicable based on the features/design of the platform.

[a] This refers to platforms that are currently centralized but have promised future integration with other federated services. The following assessment only considers their centralized trust and safety capabilities. Whether these capabilities translate into a decentralized context remains unknown.

Disparities in trust and safety capabilities between centralized and decentralized platforms can be attributed to platforms' maturity and the size of the user base. While few services make specifics about their costs and expenditures on moderation available, resourcing is a clear limiting factor, with more acute impacts on noncommercialized federated services such as Mastodon and Pixelfed. Absent the financial support that goes along with centralized, corporate social media, few parts of noncommercialized federated platforms have been able to successfully marshall the human and technological resources required to successfully execute proactive, accurate content moderation at scale. Especially with regard to persistent, high-volume malign activity, such as state-backed information operations that can be responsible for tens or even hundreds of thousands of fake accounts per month (Nimmo, Hubert, and Cheng 2021; Butler and Taege 2023), a lack of resources dedicated to ongoing monitoring turns these moderation needs into nearly insurmountable challenges for existing federated platforms.

Many of these shortcomings are solvable product and engineering challenges—and no doubt the moderation tools built into Mastodon and other federated products will improve over time. But moderation doesn't just require tooling; it also requires ongoing, sustained investment and monitoring. Heuristics that are viable one day can become inaccurate the next. Machine learning models exhibit drift over time and can either under- or over-detect the target activity. "Set it and forget it" is not a viable strategy for dealing with dedicated adversaries, and the responsible deployment of even sophisticated technical enforcement capabilities requires ongoing, sustained effort by moderators—resource-intensive capacities that interviews with platform maintainers, moderators, and researchers revealed are broadly lacking from present federated and decentralized platforms.

5 Community perspectives on moderation improvements

To elicit and enumerate potential solutions to current shortcomings in the moderation capabilities of federated and decentralized platforms, we conducted three long-form semi-structured group interviews exploring collective security threats on federated services. The interviews gathered 32 interdisciplinary experts, including developers, maintainers, administrators, and moderators of federated platforms, current and former trust and safety employees of centralized and decentralized social media companies, and academic and civil society researchers to explore possible mitigations to these threats. We solicited input on mitigations for a range of threats, including coordinated inauthentic behavior, CSAM, spam, and more. A full list of interview questions can be found in Appendix A. Interview participants highlighted a series of ongoing and possible interventions for addressing these threats, informing our recommendations with their expertise.

On the basis of these discussions, we identified three thematic areas of recommended investment to develop key functions for addressing risks on federated platforms: community, technologies, and institutions. Individually, investment in each would represent a meaningful step forward in capabilities to respond to collective security threats; jointly, these measures could significantly improve the resilience of federated services (and federated platforms broadly) to coordinated manipulation and other user-impacting harms—and create a foundational approach to threat defense that would make it easier for new entrants into the space of social media and online community to address these challenges before they take root.

5.1 Community: empower users and moderators

The moderation of federated and decentralized platforms is fundamentally a community effort, involving contributions from end users, volunteer moderators, and—sometimes—paid or professional administrators. A key component of robust governance and moderation for federated social media is the development of the necessary infrastructure for moderators and community members to be able to collaboratively identify and mitigate shared threats. Key aspects of this work include improvements to reporting capabilities on federated services, expanded use of shared denylists and community-sourced moderation labels, and more granular tools for managing the boundaries of federated networks.

Existing structures for collaborative engagement between users, moderators, and instances are largely relegated to reporting functions, and interviewees stressed that even those functions are limited and rudimentary. One interviewee reflected that improving reporting capabilities, such as allowing users to report sets of accounts promoting certain URLs, media, and hashtags, could expand user capability in detecting and flagging content for instance administrators and moderators.

In addition, shareable or centralized denylists—that is, lists of instances believed to be malicious or harmful that can be blocked en masse by instance administrators and moderators—are a useful first step for knowledge-sharing among community members, while alleviating burdens on moderators to curate and block instances individually. Initial implementations of shared instance denylists could readily extend to a critical gap identified in our analysis: an inability to exchange content moderation decisions and threat information across instance boundaries.

One interview participant suggested shareable allowlists as another alternative, where one instance can establish what they consider to be a minimally viable moderation pattern. Other instances could then be invited to join the allowlist if they agree to adhere to those terms. This provides moderators another way to set standards for who they federate with. Several participants emphasized that these tools should be easy to use and understand, to facilitate their use by moderators with limited technical skills.

Two interview participants involved in the development of federated moderation tools suggested extending these capabilities to users themselves, rather than limiting information exchange to administrators and moderators. Bluesky's model of composable moderation offers one such implementation, by allowing community members to directly contribute labels of perceived-harmful content and accounts; other users, in turn, can subscribe to those labels and apply them to filter or manage content they see (or don't see) across the Bluesky service (Graber 2023). This opens up possibilities for communities to explore different forms of moderation, and enables new forms of competition (Masnick 2023).

Effective deployment of such capabilities requires appropriate governance, oversight, and community support to enable responsible usage. In developing community-sourced denylists, two interviewees reflected that equity on an international scale is a critical factor. Major social media platforms have struggled to compile information on political bad actors from specific countries that they do not have well-defined relationships with. Community-sourced denylists are also likely to suffer the same problem. One interviewee noted that these denylists also run the risk of acting as a double-edged sword, as bad actors can co-opt those distribution techniques to popularize lists of marginalized groups on a platform that they want to silence.

Several interview participants noted that these structures would be most effective if paired with more robust capabilities for moderators and administrators to manage the interoperation of their instance with others on the network. Commonly proposed suggestions include making instances invite-only (see, for example, ClearlyClaire (2023)), or requiring some kind of trusted referral model for new sign-ups—an approach notably adopted by Bluesky during its months-long, ongoing "beta." This may well be a viable solution for parts of federated platforms that intentionally prioritize small community size and affinity based on identity or interest.

But others pointed out that the "gated community" model has at least three key challenges as a broader strategy: First, this only solves the problem of "local" manipulation, not the impacts of federated behavior on nonlocal viewers of that content. Second, it's not clear that this is actually a way to address the most sophisticated and insidious forms of manipulative behavior. Elaborately constructed inauthentic profiles—like the deep, cross-platform persona development tactics employed by an Israeli disinformation purveyor (Kirchgaessner et al. 2023)—will often withstand anything but the most invasive forms of validation. Inevitably, the more invasive validation becomes, the less usable a service is by vulnerable people and groups, who might have very good reasons for not wanting to disclose their personal information to instance operators they don't know or trust. And finally, most fundamentally, for people looking to federated platforms as an alternative to centralized social platforms like X, raising barriers to entry introduces fundamental tradeoffs against the very network effects that could help make an upstart service into a mass-market product (McArdle 2023).

An alternative structure that could strike a balance across these considerations, proposed by one interview participant, is a system of federation requests, in which instances are closed, but moderators and administrators are able to see who wishes to federate with their instances. Granular control models could help empower instance administrators and community members to intentionally interoperate with instances that adhere to agreed-upon safety and moderation standards, rather than defaulting to a more vulnerable, fully open stance.

A potentially irreconcilable tension emerges here tied to the governance of these shared moderation resources: While these proposals contemplate decentralized authority at the protocol level, and emphasize opt-in engagement, it's difficult to design a structure that would both enable auditing and validation of the legitimacy of enforcement rules, while also preventing bad actors from immediately becoming aware that they've been caught (and more critically, how they've been caught). The information that would allow one moderator to validate and agree with another's decision is the same knowledge a threat actor could use to adapt their behavior to circumvent such detections in the future. Protecting an information-sharing system from the very targets of that system requires at least some degree of centralized control and access management, but in the context of services inherently distrustful of centralized authority, it will not be easy to find a path to consensus about the governance of these systems.

5.2 Technology: make scaled enforcement effective and collective

High-level community-building provides a structure for moderation beyond the work of individual instance moderators. To effectively deploy the capabilities of a community of moderators, federated services need to develop technical capabilities for addressing threats at scale, rather than as piecemeal content moderation issues.

Most large platforms employ a two-pronged approach to moderation: First, human moderators are responsible for adjudicating individual cases about posts or accounts; second, automated systems—either driven by heuristics or machine learning—apply policy designations at scale. Depending on their technological investments, staffing, and tolerance for errors, companies might choose to allocate work across these methods in different ways; however, after a certain point of growth in usage, it's impractical to rely solely on human moderation to address harmful conduct.

An almost universal perspective shared by interviewees was an assessment that the primary gap in the moderation capabilities of federated systems is a lack of tools for automated enforcement, with manual, human moderation used as a substitute. Take, for example, the moderation of CSAM—one of the domains of trust and safety work with the greatest established body of technology and process knowledge. Currently, on platforms like Mastodon, this work is conducted manually by moderators, who have to assess content against instance-specific guidelines, download violative content to their local computers, upload it to the National Center for Missing & Exploited Children (NCMEC)'s database, then hold onto the offending material for 90 days before deleting it. In contrast, the hashing-and-matching approaches employed by centralized platforms to address CSAM and terrorism automate many of these processes, and standardize the detection of known-harmful content across platforms. Centralized platforms have long relied on proprietary hash-sharing technology such as PhotoDNA, and the hash-sharing database of harmful content maintained by the Global Internet Forum to Counter Terrorism (GIFCT) (Microsoft, n.d.; GIFCT, n.d.).

These approaches, while commonplace at large centralized platforms, are not without their own flaws and challenges. With more than 2.3 million hashes in the GIFCT database (as of 2021) (GIFCT 2019), smaller platforms considering adoption of these systems will have to either choose to employ the hash values at face value and accept the risk of false positives against their policies, or deploy significant resources to make manual moderation decisions on the basis of matches. Furthermore, as many of these algorithms are proprietary and offered for profit via APIs, those who use them need to send all media to a third-party provider—introducing privacy and governance challenges, as well as resourcing concerns for volunteer-managed instances. And even beyond the costs of procuring services from commercial vendors, the technical implementation costs of these systems can be considerable, even where the underlying systems and datasets are open-source (Farid 2021). In their discussion of CSAM challenges on federated platforms, Thiel and DiResta (2023) attribute the complicating factor to platform architecture. Federated platforms are hosted across numerous servers, each of which have their own reporting processes. When CSAM content appears on a federated service, multiple servers go through the same process of sending the same image to PhotoDNA. Multiply this across hundreds and thousands of servers, and the process becomes both costly and duplicative. Deploying these techniques for other content types would undoubtedly have similar challenges.

Addressing the technical gaps in federated moderation will require investment in specific technologies that give moderators and administrators straightforward capabilities to build, test, and deploy automated moderation. One interviewee cited the example of Twitter, where the primary system responsible for this work was called Botmaker—the core components of which were (1) a rules engine with simple syntax for constructing heuristics, (2) methods for real-time and near-real-time processing of those rules against the stream of activity on Twitter, and (3) the ability to take automated content and account moderation action based on detected rule violations (X Engineering 2014). Subsequent iterations of this system added code review and management, measurement and anomaly detection, and more robust logging and tracking of automated actions. Another interviewee spoke to how similar systems have been developed by other companies and platforms, including the Smyte platform (acquired by Twitter and subsequently shut down), and efforts from Meta and Discord (SQRL Documentation, n.d.; Stein, Chen, and Mangla 2011; Discord 2022). The exact technical design of a system for automated moderation is beyond the scope of this article, but open-source development of these capabilities would represent a meaningful contribution to the security of federated platforms.

Underscoring the importance of developing these systems, one interviewee pointed to how existing platforms with decentralized governance, like Wikipedia and OpenStreetMap, have successfully leveraged these kinds of techniques to combat abuse and spam, and could be illustrative examples for implementations on federated platforms (Nasaw 2012). Chaput (2023) has outlined how such tooling could exist on platforms such as Mastodon, which could support instance administrators in configuring multiple external trusted providers for moderated activities. Similar proposals have been advanced by the developers of Bluesky, who suggest unbundling moderation services from hosting services, to enable more flexible adoption of moderation technologies by different segments of the Bluesky network (mkantzer 2023).

Critically, interview participants emphasized that these systems should include native capabilities for sharing heuristics, indicators of compromise, and other threat intelligence about malicious actors—analogous to shared denylists employed by individual users to address unwanted interactions on social media. It is unreasonable and inefficient to expect individual instance moderators to tackle every emergent threat anew. Instead, by standardizing the syntax and structure of rules-based automated enforcement, moderators could exchange solutions to persistent problems, creating a standard set of solutions and reducing individual burden on specific instances.

5.3 Institutions: centralize response and mitigation efforts to distributed threats

The majority of developers, maintainers, and instance administrators of federated platforms we interviewed agreed that some degree of centralization is required for more effective moderation—but substantial disagreements persist about the specifics of an institutionalized approach to moderation. Nevertheless, some points of consensus emerged, with interviewees highlighting that an institutional solution to moderation on federated platforms should include (1) mechanisms for sharing technical resources and capabilities, (2) coordination mechanisms for common elements of content moderation processes, (3) measures to coordinate transparency reporting across platforms and instances, and (4) data privacy safeguards.

5.3.1 Centralized resources

Arguably the most significant constraint on federated trust and safety efforts, especially with regard to large-scale manipulative behavior, is that no one has access to platform-wide data, and therefore cannot conduct analysis of coordinated threats across servers on federated services.

Detection of behavioral manipulation is in large part reliant on access to data about on-platform activity—and the openness of federated platforms has largely resulted in the ready availability of APIs to enable this kind of access. For example, Mastodon has a robust set of public APIs that would allow researchers to gain authenticated access to real-time data about conversations happening on a given instance (Mastodon, n.d.). But interview participants also pointed to how federation complicates the use of these APIs to study ecosystem-level threats: Whereas Twitter's APIs offer a single channel for collecting data about all the activity happening globally across the Twitter service, Mastodon's APIs are mostly instance-specific. As a result, many data-collection efforts involve just focusing on a handful of the largest instances (instances.social, n.d.). Alternatively, researchers seeking a network-wide perspective have to devise novel datagathering strategies that involve collecting data from successively smaller and smaller instances—methods that create the potential for inconsistency and gaps in data, and that are time- and resource-intensive. While other federated platforms, like Bluesky, offer more direct, public access to a global, consolidated firehose (at least theoretically spanning all instances of the service), interview participants indicated that the differences in network structure across federated platforms makes standardized data collection challenging.

More concerningly, the challenges posed by a lack of network-wide data are not exclusive to external researchers; moderators and administrators of federated instances themselves face similar obstacles to meaningful threat identification. Many of the techniques employed by large platforms to detect manipulation involve surveying the full population of accounts and activity, and looking for unusual clusters or patterns of behavior within that population—a practice of threat identification using centralized telemetry. In federated systems, instance administrators only have comprehensive logs for the activity of local users of their particular instance. Interview participants noted that a threat actor who spreads their inauthentic accounts across a handful of the biggest instances is both less likely to be caught as behaviorally anomalous by any one instance, and less likely to have the full scope of their operation, across all the instances on which they operate, detected.

Federated platforms must also contend with the challenges of moderating distributed but coordinated threats. Mastodon's moderation capabilities, for example, provide for a few rudimentary antispam techniques for addressing scaled threats (which even the Mastodon documentation notes will be circumvented by dedicated spammers (Mastodon, n.d.)—but interview participants indicated that largely, Mastodon moderation is focused on either dealing with individually problematic users (by restricting or banning them from a given instance), or defederating an entire instance (Mastodon, n.d.). Spam and platform manipulation are unlikely to be solvable using this tactic, because they primarily manifest as distributed threats across mainstream, non-malicious instances. Put another way, sophisticated adversarial threats will not concentrate themselves on single instances, waiting to be defederated. Instead, inauthentic accounts are likely to be dispersed across mainstream servers. Interview participants stressed that this dispersion creates a distributed burden of detection across already overworked and under-equipped instance moderators, who have to deal with these accounts one by one, instance by instance.

Most participants agree that centralizing resources for threat hunting could mitigate some of these challenges. This includes both (1) access to network-level data and (2) financial support.

A core challenge repeatedly emphasized by interviewees is the lack of a unified evidence base for threat hunting. An institutionalized solution would require participating instances to pool their data for shared analysis. In place of hundreds or thousands of separate repositories, a central data source could enable the kind of aggregate analysis and anomaly detection that represents the core of counter-manipulation efforts. This pooled data could also let moderators and administrators identify whether an account has multiple iterations across instances. This could be particularly useful for detecting coordinated inauthentic behavior from state actors or sophisticated operators, which often have identical content, behavioral patterns, and account specifics.

Several interview participants emphasized that establishing a sustainable model of financial support is a key part of ensuring this work is viable as federated platforms scale. One solution, proposed by several interviewees, was to establish collective response efforts as nonprofit ventures, independent of specific platforms or service providers operating in the safety and security space. While interviewees disagreed about and debated the merits of various funding structures, options suggested include a mixture of contributions by participating instances, users, and nongovernmental, independent funding sources. Chiefly, these resources could fund the storage and compute costs tied to analysis performed by participating analysts. A core staff of paid analysts could help accelerate the development of the necessary systems expertise and longitudinal threat awareness in this space.

5.3.2 Coordination mechanisms

A critical gap identified in our interviews and discussions with moderators and administrators is an inability to coordinate trust and safety work with other instances and services facing shared threats. Interviewees noted that such sharing efforts would enable them to build on knowledge and experience about this work that already exists at other companies or platforms, to avoid having to reinvent standard or recurrent moderation solutions. While “coordination” can take a wide range of forms—at least some of which, like top-down imposition of moderation decisions, could be antithetical to the core ethos of decentralization—interview participants identified three areas where centralized coordination mechanisms could play a meaningful role: (1) training and support for moderators, (2) user recourse mechanisms, and (3) representation of federated platforms in broader industry and governmental discussions.

A recurring challenge cited by instance operators and moderators is that many lack the time, knowledge, tools, governance frameworks, and inclination necessary to do the highly specialized work of disinformation detection and analysis. Staff responsible for this work at centralized platforms reported in our interview that training programs to get even technically proficient analysts fully up to speed on advanced analytic techniques can take months. A centralized institution could provide training and resources on how to detect and address threats, and to account for the community’s evolving needs. A key element of this is developing formalized channels for communication among moderators across instances; a centralized institution could serve a critical convening function and enable this type of communication.

Even as interviewees generally agreed that some form of cross-instance investigation and enforcement is essential for effective governance, several interviewees noted that these practices necessitate some degree of coordination around the experiences of users impacted by these processes. Federated platforms by their nature rely on highly distributed user-to-platform relationships, in which there is not a single, service-wide party responsible for user experience. However, errors are an inevitable part of content moderation at scale—and even an effectively operating institution will make analytic mistakes that have direct consequences on the users of federated services. These errors, and the steps required to investigate and reverse them, could be considered the responsibility of the centralized institution—not individual instance operators, who would not have full knowledge of the reasons or evidence behind a moderation action impacting their users.

Finally, a centralized institution can act as a coordination mechanism for representatives across centralized and decentralized platforms. Interviewees noted that it can be challenging if not impossible to recognize individual accounts or posts as connected to a disinformation campaign in the absence of cross-platform awareness of related conduct. The largest platforms—chiefly, Meta, Google, and Twitter (pre-acquisition)—regularly shared information, including specific indicators of compromise tied to particular campaigns, with other companies in the ecosystem in furtherance of collective security (Gleicher et al. 2021; Shields 2024). Information-sharing among platform teams represents a critical way to build this awareness—and take advantage of gaps in adversaries' operational security to detect additional deceptive accounts and campaigns.

Several interviewees remarked that federation by its very nature makes this kind of cross-platform collaboration difficult. Thousands of individual instance operators each have responsibility for a potential target of this conduct, but it's infeasible for larger platforms, like Meta and Google, to engage with moderators or administrators from each instance directly. Even assuming platforms limit these engagements to only a handful of the largest instances across federated services, the legal frameworks and contractual protections needed to share data across platforms without running afoul of privacy regulations like the General Data Protection Regulation (GDPR) require specialized legal expertise and negotiation, which is often out of reach for hobbyist efforts. In addition, absent an institutionalized way to verify the trustworthiness and legitimacy of instance administrators and moderators, larger platforms will have limited information about who they are working with—and correspondingly may either choose not to engage, or will feel constrained in their ability to share relevant data. They may be concerned that bad actors posing as moderators of legitimate instances on federated platforms could leverage these structural ambiguities to gain access to larger platforms' staff and intel, creating commercial, political, and privacy risk.

Driven by these factors, interview participants stressed the need for federated platforms to have representation within the organizations and working groups responsible for this work industry-wide. Commonly cited examples of these groups include professional organizations like the Trust and Safety Professionals Association (TSPA), industry self-regulatory and standards bodies like the Digital Trust and Safety Partnership (DTSP), and specialized groups like the Tech Coalition (an organization focused on mitigating online child exploitation) and GIFCT. Several interview participants with experience working on counter-manipulation efforts at tech platforms also noted that engagement with law enforcement and intelligence agencies would be valuable—although such efforts are presently complicated, especially in the United States, by broader debates about government "jawboning" and censorship (Nix and Zakrzewski 2023).

5.3.3 Transparency

Interviewees stressed that any centralized institution (or institutions) operating in this space should prioritize clear and exhaustive transparency reporting, including at minimum the routine reporting of aggregated statistics about an institution's operations and actions. In the context of federated services, transparency reporting could include data about moderation actions taken, individuals and groups that have been designated as harmful, and the policies and standards used to make moderation determinations (Lai, Shiffman, and Wanless 2023). Developing these capabilities centrally could help fill a broader transparency gap impacting federated platforms, enabling community sense-making about platform norms and providing users agency in their selection of instances and moderation approaches. This remains a critical gap at present, where users generally have little insight into or understanding of how different instances operate, beyond word-of-mouth.

Several interviewees argued that having each individual instance conduct transparency reporting would not be a tenable approach. The first challenge lies in resourcing: Instance moderators are unpaid, volunteer individuals, and do not have the time, training, or bandwidth to conduct the data collection and analysis required for transparency reporting (on top of and beyond the already substantial burdens of actually conducting moderation). Additionally, moderators of small or controversial instances that conduct transparency reporting may risk deanonymization and harassment by those that do not agree with their decisions. Several interviewees who advocated for institutional moderation solutions noted that these institutions could centralize transparency reporting about their efforts, and in the process help alleviate these burdens on instance moderators by placing the responsibility on a larger, more well-resourced body. Cumulative reporting also provides a layer of protection and anonymity to moderators.

5.3.4 Data privacy safeguards and standards

Centralized telemetry by its very nature involves the collection, aggregation, and analysis of sensitive user data—and interview participants emphasized the need to design systems for centralized trust and safety work that appropriately balance privacy and safety equities.

Specifically, the types of log and user data necessary for detecting and counteracting manipulative behavior are often high-risk, cornerstone data from the perspective of users and service providers. In some discussions about Mastodon moderation, for example, administrators have said they resist looking at the information they could access through an instance's administrative tools, citing privacy concerns—and a repository containing logs from across federated platforms becomes an obvious and appealing intrusion target. The privacy/security trade-offs here cannot be avoided, but they can be mitigated by investments in appropriate protocols for data management and sharing. Elements of a viable approach could include technical safeguards on information security (such as access management and credentialing for sensitive data types), hash-based approaches for exchanging user- or device-specific information without exposing raw identifiers (like email addresses or phone numbers), administrator tools with built-in restrictions that save and log administrator activity, and secure compute environments for analyzing cross-platform data without the option of exporting or storing it.[4] Interviewees also suggested the use of legal mechanisms, such as developing legal documentation that would hold moderators liable for looking up other users' personal data through administrator tools.

Critically, centralized moderation solutions focused on spam, platform manipulation, and other behavioral threats should be kept separate from other moderation tasks and challenges—even if similar structures might be useful in those contexts as well. While the line between content-driven issues like misinformation and behavioral- and actor-based challenges can be blurry, maintaining boundaries between security and manipulation interventions and broader content moderation can help build trust that this work is rooted in addressing collective security risks, not carrying out ideological censorship.

6 Discussion

Despite the challenges and shortcomings in moderation capabilities identified in our analysis, it's worth recognizing that federated services, at least in their present implementations, have some inherent resilience to some of the most common social media threats, including platform manipulation and spam. None of the existing federated platforms have implemented algorithmic content recommendations as a key component of user experience, resulting in a smaller attack surface for inauthentic engagement and behavioral manipulation. And while Mastodon has introduced a version of a "trending topics" list, such features tend to rely on aggregation of local (rather than global or federated) activity, which removes much of the incentive for engaging in large-scale spam.

[4] One specific example of a suitable technology, cited by an interviewee, are so-called "data capsules," which are virtual machines that allow flexible computation but are designed to prevent the exfiltration of content, requiring human review before any analysis is released. See, for example, HTRC Analytics (n.d.).

The lack of built-in monetization programs on virtually all federated platforms—at least presently—likewise reduces incentives for programmatic malfeasance.

However, in most other aspects, federated services remain vulnerable to many of the same threats that have plagued centralized platforms for years. To address these risks, we draw from participant input and our own analysis to surface four key recommendations toward enhancing collective security on federated services:

6.1 Institutionalize shared responses to critical harms

Our findings show that decentralized moderation is challenging primarily because each instance operator has to reinvent many of the policies and procedures of moderation for themselves. As developer Werdmuller (2022) puts it, "While software is provided to technically moderate, there are very few ecosystem resources to explain how to approach this from a human perspective." The results are predictable for anyone familiar with the challenges of social media content moderation: Users report erroneous or inexplicable bans, with limited recourse from volunteer administrators moonlighting as content moderators (Masnick 2019b). Larger-scale harassment campaigns overwhelm victims and administrators alike (Sockwell 2018).

Recognizing these challenges, we endorse a path forward that would institutionalize and centralize some critical moderation functions. Many of the solutions proposed by interviewees in our study involve some degree of centralization—a concept that seems at odds with federation's inherent focus on decentralized approaches to development and organization. The suggestion of centralization is certainly controversial among proponents of federation, drawing from differences in ideology and disagreements over how meaningful governance challenges could be solved (xg15 2022; h310s 2023). However, the suggestion is not wholly anathema, with many moderators and administrators signaling a reasonably high desire for some form of centralized solution to the challenges they face (IFTAS 2023). In a way, centralization versus decentralization in the context of threat response is less a question of whether to create a centralized solution to these problems, and more a matter of what kind of centralization is preferred by community members. Writing about the labor and economic challenges of moderation on federated platforms, Rozenshtein (2022) draws a comparison to email, noting that,

> If effective moderation turns out to require more infrastructure, that could lead to a greater consolidation of instances. This is what happened with email, which, in part due to the investments necessary to counter spam, has become increasingly dominated by Google and Microsoft. If similar scale is necessary to fight spam and bot accounts, this could serve as a centripetal force to counter the Fediverse's decentralized architecture and lead to a Fediverse that is more centralized than it is today [...] .

We agree with Rozenshtein's basic intuition that infrastructural costs—including, critically, the cost of labor for human moderation—will force certain trade-offs against full decentralization; we do not see a future state where present moderation models continue to keep pace with user growth of services like Mastodon and Bluesky. However, we are less certain that the centralization and consolidation of instances is preferable to the centralization of certain, particular moderation functions in trusted intermediaries. Spam, coordinated manipulation, and other behavioral threats are sufficiently different in kind from other types of content-focused moderation to uniquely require and benefit from this kind of centralization, and we believe that investment in such a centralized resource need not undermine the core tenets of decentralized, federated social media.

6.2 Build transparent governance into the system

At the core of governance in the context of trust and safety is the “who watches the watchers?” dilemma. These challenges are especially acute in the context of decentralized systems designed and built around an apprehension toward unaccountable centralized authorities. Independent expert oversight therefore becomes an essential property of a functional system, especially when the reasons for a moderation decision are not immediately legible to lay observers (as is the case with many types of behavioral manipulation). Building an institution that operates from the start with transparency and accountability to independent observers could help alleviate wariness about outsourcing moderation tasks to any centralized entity. Academic researchers are likely the most credibly positioned to monitor for and report on the activities of collective moderation efforts; appropriate data, including comprehensive archives of moderated content and accounts, could be available for these purposes.

In keeping with the decentralized ethos of federated platforms’ development, even a centralized approach to threat prevention could take the form of a structured hub for activity, not a monolithic solution. Much of the most effective investigation of social media manipulation (chiefly on Twitter, to date) has originated with hobbyist, academic, and civil society groups; a collective structure could focus on supporting and enabling these efforts. Appropriately experienced practitioners with an interest in contributing should be allowed to do so, regardless of their institutional affiliation. Keeping bad actors out will require ongoing scrutiny and governance, including the development of standards for who has access to the data and tools connected with collective response efforts; this governance could be rules-based and subject to transparency reporting and oversight.

6.3 Invest in tooling

The challenge of establishing sustainable funding structures looms large over efforts to develop and implement scalable tooling for federated moderation. There are many existing efforts to improve tooling, all of which cannot be implemented without sufficient funding. In the short term, increased funding for tool development can help address these limitations. Funding in the space needs to account for the needs and norms of federated services, such as being open to applications not just from companies but also individuals.

Longer-term investment in tooling can be supported by a centralized institution that curates resources, maintains tools, and connects tool builders with funders that would like to invest in this space. Such an organization answers calls to better institutionalize the field of trust and safety, which have become increasingly commonplace (Slater and Masiello 2023). Adapting to the needs of decentralized platforms, such efforts need to allow for the building and maintaining of critical trust and safety technologies in public, outside of existing large, centralized platforms.

6.4 Enable data sharing across instances

Finally, any effective approach to federated moderation must include tooling that supports data sharing across instances. One proposal in circulation calls for a “Fediverse Moderation Tool” system that would leverage the same ActivityPub protocol powering federated platforms to allow moderators to share “abuse intelligence” about their actions (ElanHasson 2023). Using this kind of system, instance moderators could opt into automatically adopting the moderation decisions of their trusted peers—reducing the need to manually ban roving abusive actors or manually curate lists of malicious domains or IPs. A similar structure could accommodate the exchange of more complex rules and heuristics for scaled enforcement. This kind of infrastructure means that instance administrators and moderators could share information share and learn from each other, instead of having to develop moderation capacities from scratch. Enabling communication on an administrator-to-administrator level also opens up avenues for accountability mechanisms. For example, administrators could develop partnership agreements to audit each others' activity and moderation decisions.

7 Conclusion

As users turn to federated alternatives to mainstream services, gaps in platform moderation capabilities will quickly become apparent, and represent a failure condition for emergent services. If federated platforms hope to become viable for more than their existing core constituencies of early adopters, the community of federated platform maintainers will need to develop solutions to consumer needs for safety and security. Ultimately, robust approaches to collective threat mitigation will be an essential part of how federated services foster mainstream adoption, and manage the adversarial targeting that goes along with it.

The present state of federated trust and safety remains nascent (IFTAS 2023). Many federated platform instances reported limited moderation coverage outside of a narrow window of hours each day. Few instance administrators represented that they understand or have structures to manage legal responsibilities associated with hosting user-generated content (Paolucci 2022). A majority of respondents to a survey of federated platform moderators reported that the instances they govern do not have any formal guidance or training for moderators. A number of moderators reported experiencing burnout as a result of their responsibilities, with nowhere to turn for support. Absent action, we should expect these challenges to scale along with federated platforms themselves.

Despite these challenges, coordinated action to address collective security risks facing federated platforms is possible—and the community of developers and practitioners engaged with these issues continues to grow. Federation presents a range of possibilities for new models of governance and community building, which can and should be deployed against social media's long-standing threats and harms. Institutionalizing shared responses to critical harms, building transparent governance into federated services, creating an iterative model for tool development, and allowing data sharing across instances are essential parts of a solution.

References

Access Now. n.d. “Transparency Reporting Index.” Accessed January 22, 2024. https://www.accessnow.org/campaign/transparency-reporting-index/.

ActivityPub. n.d. “ActivityPub.” Accessed January 18, 2024. https://activitypub.rocks/.

Arif, Ahmer, Leo Graiden Stewart, and Kate Starbird. 2018. “Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse.” In *Proceedings of the ACM on Human-Computer Interaction,* 2:1–27. CSCW. November 1, 2018. https://doi.org/10.1145/3274289.

Bluesky. n.d. “Bluesky.” Accessed January 18, 2024. https://blueskyweb.xyz/.

Bradshaw, Samantha, Ualan Campbell-Smith, Amelie Henle, Antonella Perini, Sivanne Shalev, Hannah Bailey, and Philip N. Howard. 2021. *Country Case Studies, Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation.* Research report. Oxford Internet Institute, January 13, 2021. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/03/Case-Studies_FINAL.pdf.

Butler, Zak, and Jonas Taege. 2023. “Over 50,000 instances of DRAGONBRIDGE activity disrupted in 2022.” *Google Threat Analysis Group* (January 26, 2023). https://blog.google/threat-analysis-group/over-50000-instances-of-dragonbridge-activity-disrupted-in-2022/.

Chancellor, Stevie, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. “#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities.” In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing,* 1201–13. February 27, 2016. https://doi.org/10.1145/2818048.2819963.

Chaput, Renaud. 2023. “Evolving Mastodon’s Trust & Safety Features.” *Renaud Chaput blog* (July 19, 2023). https://renchap.com/blog/post/evolving_mastodon_trust_and_safety/.

Chen, Adrian. 2015. “The Agency.” *The New York Times Magazine* (June 2, 2015). https://www.nytimes.com/2015/06/07/magazine/the-agency.html.

Chen, Jesse. 2023. “Threads: The inside story of Meta’s newest social app.” *Engineering at Meta* (September 7, 2023). https://engineering.fb.com/2023/09/07/culture/threads-inside-story-metas-newest-social-app/.

ClearlyClaire. 2023. “Add support for server-specific emergency rules,” May 18, 2023. https://github.com/mastodon/mastodon/pull/25032.

Conlon, Patrick, William Nuland, and Kanishk Karan. 2022. “Investigating Influence Operations by Twitter Integrity.” In *Perspectives for Influence Operations Investigators,* edited by Victoria Smith, Jon Bateman, and Dean Jackson. Carnegie Endowment for International Peace, October 25, 2022.

Cramer, Geoffrey, William P. Maxam III, Qianqian Li, and James C. Davis. 2023. “An Exploratory Empirical Study of Trust & Safety Engineering in Social Media Platforms,” https://davisjam.github.io/files/publications/CramerMaxamLiDavis-TrustAndSafetyEngineeringInSMPs.pdf.

DFRLab. 2019. “#InfluenceForSale: Venezuela’s Twitter Propaganda Mill.” *Medium* (February 3, 2019). https://medium.com/dfrlab/influenceforsale-venezuelas-twitter-propaganda-mill-cd20ee4b33d8.

Discord. 2022. “Auto Moderation In Discord,” June 3, 2022. https://discord.com/safety/auto-moderation-in-discord.

Donovan, Joan, and Brian Friedberg. 2019. *Source Hacking: Media Manipulation in Practice.* Research report. Data & Society, September 4, 2019. https://datasociety.net/library/source-hacking-media-manipulation-in-practice/.

Duffy, Clare. 2023. “Threads now has ‘tens of millions’ of daily users. But its honeymoon phase may be over.” *CNN* (July 19, 2023). https://www.cnn.com/2023/07/19/tech/threads-meta-growth-plan/index.html.

ElanHasson. 2023. “Fediverse Moderation Tools Proposal,” October 25, 2023. https://gist.github.com/ElanHasson/f2212425a459964a1210bda50806236d.

Ermoshina, Ksenia, and Francesca Musiani. 2022. “Safer spaces by design? Federated architectures and alternative socio-technical models for content moderation.” *Annual Symposium of the Global Internet Governance Academic Network (GigaNet)* (November). https://hal.science/hal-03930548/.

European Commission. 2022. *2022 Strengthened Code of Practice on Disinformation,* June 16, 2022. https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

European Union. 2022. “Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance),” no. PE/30/2022/REV/1 (October 19, 2022). https://eur-lex.europa.eu/eli/reg/2022/2065/oj.

Farid, Hany. 2021. “An Overview of Perceptual Hashing.” *Journal of Online Trust and Safety* 1, no. 1 (October 28, 2021). https://doi.org/10.54501/jots.v1i1.24.

FediSeer. n.d. “Mastodon.” Accessed January 18, 2024. https://fediseer.com/.

François, Camille. 2019. “Actors, Behaviors, Content: A Disinformation ABC.” In *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression,* edited by Susan Ness, Marietje Schaake, and Kathleen Hall Jamieson. September 20, 2019. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/ABC_Framework_TWG_Francois_Sept_2019.pdf.

Gehl, Robert W., and Diana Zulli. 2022. “The digital covenant: non-centralized platform governance on the Mastodon social network.” *Information, Communication & Society* 26 (16 2022): 3275–91. https://doi.org/10.1080/1369118X.2022.2147400.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press, June 26, 2018.

Giridharadas, Anand. 2022. “No One Wants A Pizzaburger.” *The Atlantic* (October 4, 2022). https://www.theatlantic.com/ideas/archive/2022/10/russia-social-media-troll-farm-persuasion-american-unity-book/671635/.

Gleicher, Nathaniel, Margarita Franklin, David Agranovich, Ben Nimmo, Olga Belogolova, and Mike Torrey. 2021. *Threat Report: The State of Influence Operations 2017–2020.* Research report. Facebook, May 20, 2021. https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf.

Global Internet Forum to Counter Terrorism. 2019. *Transparency Report 2021.* Research report. GIFCT, July 25, 2019. https://gifct.org/wp-content/uploads/2021/07/GIFCT-TransparencyReport2021.pdf.

Global Internet Forum to Counter Terrorism. n.d. “GIFCT’s Hash-Sharing Database.” Accessed January 18, 2024. https://gifct.org/hsdb/.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. “Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance.” *Big Data & Society* 7, no. 1 (February 28, 2020). https://doi.org/10.1177/2053951719897945.

Graber, Jay. 2020. “Decentralized Social Networks.” *Medium* (January 9, 2020). https://medium.com/decentralized-web/decentralized-social-networks-e5a7a2603f53.

———. 2023. “Composable Moderation.” *Bluesky Blog* (April 13, 2023). https://blueskyweb.xyz/blog/4-13-2023-moderation.

Graphika and Stanford Internet Observatory. 2023. *Bad Reputation: Suspected Russian Actors Leverage Alternative Tech Platforms in Continued Effort to Covertly Influence Right-Wing U.S. Audiences.* Technical report. Stanford Digital Repository, April 21, 2023. https://purl.stanford.edu/pm393tq4393.

Graphika Team. 2020. *Step into My Parler.* Research report. Graphika, October 1, 2020. https://public-assets.graphika.com/reports/graphika_report_step_into_my_parler.pdf.

Griffin, Rachel. 2023. “Algorithmic Content Moderation Brings New Opportunities and Risks.” *Centre for International Governance Innovation* (October 23, 2023). https://www.cigionline.org/articles/algorithmic-content-moderation-brings-new-opportunities-and-risks/.

h310s. 2023. “The ‘Hidden Dangers’ of the Decentralized Web: in other words, not trusting centralized corporate platforms leads to right-wing extremism/anti semitic conspiracy theories. C’mon,” May 19, 2023. Accessed January 23, 2024. https://www.reddit.com/r/Mastodon/comments/13m2et9/the_hidden_dangers_of_the_decentralized_web_in/.

HTRC Analytics. n.d. “Data Capsules.” Accessed January 24, 2024. https://analytics.hathitrust.org/staticcapsules.

Independent Federated Trust and Safety. 2023. *IFTAS Fediverse Moderator Needs Assessment Results.* Research report. IFTAS, October 10, 2023. https://drive.google.com/file/d/1gtAzAdv_21W50csbrdBe1ORl2wtQ9pDs/.

Infantino, Susan. 2013. “Transparency Report: Government removal requests continue to rise.” *Google* (December 19, 2013). https://blog.google/outreach-initiatives/public-policy/transparency-report-government-removal/.

instances.social. n.d. “Mastodon instances.” Accessed January 18, 2024. https://instances.social/list/advanced#lang=&allowed=&prohibited=&min-users=&max-users=.

Keller, Daphne. 2021. “The Future of Platform Power: Making Middleware Work.” *Journal of Democracy* 32 (3): 168–72. https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/.

Keller, Daphne, and Paddy Leerssen. 2020. “Facts and where to find them: Empirical research on internet platforms and content moderation.” In *Social Media and Democracy: The State of the Field and Prospects for Reform,* edited by Nathaniel Persily and Joshua Tucker, 220–51. Cambridge University Press, July 9, 2020. https://ssrn.com/abstract=3504930.

Kirchgaessner, Stephanie, Manisha Ganguly, David Pegg, Carole Cadwalladr, and Jason Burke. 2023. “Revealed: the hacking and disinformation team meddling in elections.” *The Guardian* (February 15, 2023). https://www.theguardian.com/world/2023/feb/15/revealed-disinformation-team-jorge-claim-meddling-elections-tal-hanan.

Klonick, Kate. 2017. “The New Governors: The People, Rules, and Processes Governing Online Speech.” *Harvard Law Review* 131 (6): 1598–670. https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/.

Komaitis, Konstantinos, and Louis-Victor De Franssu. 2022. “Can Mastodon Survive Europe’s Digital Services Act?” *Tech Policy Press* (November 16, 2022). https://www.techpolicy.press/can-mastodon-survive-europes-digital-services-act/.

Koutrika, Georgia, Frans Adjie Effendi, Zolt’n Gyöngyi, Paul Heymann, and Hector Garcia-Molina. 2008. “Combating spam in tagging systems: An evaluation.” *ACM Transactions on the Web* 2, no. 22 (4 2008): 1–34. https://doi.org/10.1145/1409220.1409225.

Lai, Samantha, Naomi Shiffman, and Alicia Wanless. 2023. “Operational Reporting By Online Services: A Proposed Framework.” *Carnegie Endowment for International Peace* (May 18, 2023). https://carnegieendowment.org/2023/05/18/operational-reporting-by-online-services-proposed-framework-pub-89776.

Levine, Brian Neil. 2022. *Increasing the Efficacy of Investigations of Online Child Sex Exploitation.* Research report. U.S. Department of Justice Office of Justice Programs, May. https://static1.squarespace.com/static/5b7ea2794cde7a79e7c00582/t/6317425ef826c84ff2a967fd/1662468705076/Increasing+the+efficiency.pdf.

Library of Congress. 2021. *Germany: Network Enforcement Act Amended to Better Fight Online Hate Speech,* July 6, 2021. https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/.

Linvill, Darren L., and Patrick L. Warren. 2020. “Troll Factories: Manufacturing Specialized Disinformation on Twitter.” *Political Communication* 37 (4 2020): 447–67. https://doi.org/10.1080/10584609.2020.1718257.

Lukito, Josephine. 2019. “Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017.” *Political Communication* 37, no. 2 (October 14, 2019): 238–55. https://doi.org/10.1080/10584609.2019.1661889.

MacCarthy, Mark. 2022. “Transparency Recommendations for Regulatory Regimes of Digital Platforms.” *Centre for International Governance Innovation* (March 8, 2022). https://www.cigionline.org/publications/transparency-recommendations-for-regulatory-regimes-of-digital-platforms/.

Mansoux, Aymeric, and Roel Roscam Abbing. 2019. “Seven Theses on the Fediverse and the Becoming of FLOSS.” In *The Eternal Network: The Ends and Becomings of Network Culture,* edited by Kristoffer Gansing and Inga Luchs, 135–40. December 20, 2019. https://monoskop.org/images/c/cc/Mansoux_Aymeric_Abbing_Roel_Roscam_2020_Seven_Theses_on_the_Fediverse_and_the_Becoming_of_FLOSS.pdf.

Masnick, Mike. 2019a. “Gab, Mastodon and the Challenges of Content Moderation on A More Distributed Social Network.” *Techdirt* (July 16, 2019). https://www.techdirt.com/2019/07/16/gab-mastodon-challenges-content-moderation-more-distributed-social-network/.

Masnick, Mike. 2019b. “Masnick’s Impossibility Theorem: Content Moderation at Scale Is Impossible to Do Well.” *Techdirt* (November 20, 2019). https://www.techdirt.com/2019/11/20/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well/.

———. 2019c. “Protocols, Not Platforms: A Technological Approach to Free Speech.” *Knight First Amendment Institute at Columbia University* (August 21, 2019). https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech.

———. 2023. “Bluesky Plans Decentralized Composable Moderation.” *Techdirt* (April 20, 2023). https://www.techdirt.com/2023/04/20/bluesky-plans-decentralized-composable-moderation/.

Mastodon. 2023. “Mastodon Annual Report 2022,” October 2, 2023. https://joinmastodon.org/reports/Mastodon%20Annual%20Report%202022.pdf.

———. n.d. “Mastodon Users.” Accessed January 19, 2024. https://mastodon.social/@mastodonusercount.

———. n.d. “Moderation actions.” Accessed January 18, 2024. https://docs.joinmastodon.org/admin/moderation/.

———. n.d. “Playing with public data.” Accessed January 18, 2024. https://docs.joinmastodon.org/client/public/.

Matias, J. Nathan. 2016. “Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,* 1138–51. May 7, 2016. https://natematias.com/media/GoingDark-Matias-2016.pdf.

McArdle, Megan. 2023. “Opinion | Twitter might be replaced, but not by Mastodon or other imitators.” *The Washington Post* (January 17, 2023). https://www.washingtonpost.com/opinions/2023/01/17/twitter-mastodon-replacement-social-media/.

Microsoft. n.d. “PhotoDNA.” Accessed January 19, 2024. https://www.microsoft.com/en-us/photodna.

mkantzer. 2023. “Bluesky Social Proposals 0002: Labeling and Moderation Controls,” June 24, 2023. https://github.com/bluesky-social/proposals/tree/main/0002-labeling-and-moderation-controls#readme.

Mohammad, Linah, Patrick Jarenwattananon, and Scott Detrow. 2023. “Threads, Meta’s competitor to Twitter, is off to a fast start.” *NPR* (July 7, 2023). https://www.npr.org/2023/07/07/1186531843/threads-metas-competitor-to-twitter-is-off-to-a-fast-start.

Nasaw, Daniel. 2012. “Meet the ‘bots’ that edit Wikipedia.” *BBC* (July 25, 2012). https://www.bbc.com/news/magazine-18892510.

Nicholson, Matthew N., Brian C. Keegan, and Casey Fiesler. 2023. “Mastodon Rules: Characterizing Formal Rules on Popular Mastodon Instances.” *CSCW ’23 Companion: Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (October 14, 2023): 86–90. https://doi.org/10.1145/3584931.3606970.

Nimmo, Ben, and David Agranovich. 2022. “Removing Coordinated Inauthentic Behavior From China and Russia.” *Meta Newsroom* (September 27, 2022). https://about.fb.com/news/2022/09/removing-coordinated-inauthentic-behavior-from-china-and-russia/.

Nimmo, Ben, C. Shawn Eib, and L. Tamora. 2019. *Cross-Platform Spam Network Targeted Hong Kong Protests.* Technical report. Graphika, September. https://public-assets.graphika.com/reports/graphika_report_spamouflage.pdf.

Nimmo, Ben, Ira Hubert, and Yang Cheng. 2021. *Spamouflage Breakout.* Technical report. Graphika, February. https://public-assets.graphika.com/reports/graphika_report_spamouflage_breakout.pdf.

Nix, Naomi, and Cat Zakrzewski. 2023. "U.S. stops helping Big Tech spot foreign meddling and GOP legal threats." *The Washington Post* (November 30, 2023). https://www.washingtonpost.com/technology/2023/11/30/biden-foreign-disinformation-social-media-election-interference/.

Ong, Jonathan Corpus, and Jason Vincent A. Cabañes. 2018. "Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines." *Communication Department Faculty Publication Series* 74. https://doi.org/10.7275/2cq4-5396.

Paolucci, Denise. 2022. "A guide to potential liability pitfalls for people running a Mastodon instance," November 20, 2022. https://denise.dreamwidth.org/91757.html.

Puddephatt, Andrew. 2021. *Letting the Sun Shine In: Transparency and Accountability in the Digital Age.* Research report. UNESCO, May 3, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000377231.

Rid, Thomas, and Ben Buchanan. 2014. "Attributing Cyber Attacks." *Journal of Strategic Studies* (December 23, 2014): 4–37. https://doi.org/10.1080/01402390.2014.977382.

Roberts, Sarah T. 2022. "Content Moderation." In *Encyclopedia of Big Data,* edited by Laurie A. Schintler and Connie L. McNeely, 211–14. February 12, 2022. https://doi.org/10.1007/978-3-319-32010-6_44.

Roberts, Sarah T., Stacy E. Wood, and Yvonne Eadon. 2023. "'We Care About the Internet; We Care About Everything' Understanding Social Media Content Moderators' Mental Models and Support Needs." In *Proceedings of the 56th Hawaii International Conference on Systems Sciences.* January 3, 2023. https://par.nsf.gov/servlets/purl/10455915.

Roth, Emma. 2023. "Threads is struggling to retain users — but it could still catch up to X." *The Verge* (September 26, 2023). https://www.theverge.com/2023/9/26/23890592/threads-meta-monthly-users-data-x-twitter.

Rozenshtein, Alan. 2022. "Moderating the Fediverse: Content Moderation on Distributed Social Media." *Journal of Free Speech Law* 3 (1 2022): 217–36. https://doi.org/10.2139/ssrn.4213674.

Shields, Wesley. 2024. "Russian threat group COLDRIVER expands its targeting of Western officials to include the use of malware." *Google Threat Analysis Group* (January 18, 2024). https://blog.google/threat-analysis-group/google-tag-coldriver-russian-phishing-malware/.

Slater, Derek, and Betsy Masiello. 2023. "Annex 2: Building Open Trust and Safety Tools." In *Scaling Trust on the Web,* edited by The Atlantic Council of the United States. June 21, 2023. https://www.atlanticcouncil.org/wp-content/uploads/2023/06/scaling-trust-on-the-web_annex2.pdf.

Sockwell, Daniel Long. 2018. "Mastodon Mobs and Mastodon Mods: Dealing with Outside Groups Pressuring Instance Administrators." *Codesections,* accessed January 19, 2024. https://www.codesections.com/blog/mastodon-mobs-and-mastodon-mobs.

Spence, Ruth, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. "The Psychological Impacts of Content Moderation on Content Moderators: A Qualitative Study." *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17, no. 4 (September 13, 2023). https://doi.org/10.5817/CP2023-4-8.

SQRL Documentation. n.d. "SQRL: A Safe, Stateful Language for Event Streams." Accessed January 19, 2024. https://sqrl-lang.github.io/sqrl/.

Starbird, Kate, Renée DiResta, and Matt DeButts. 2023. "Influence and Improvisation: Participatory Disinformation during the 2020 US Election." *Social Media + Society* 9, no. 2 (June 7, 2023). https://doi.org/10.1177/20563051231177943.

Stein, Tao, Roger Chen, and Karan Mangla. 2011. *Facebook Immune System.* Research report. Facebook Research, April 10, 2011. https://research.facebook.com/publications/facebook-immune-system/.

Struett, Thomas, Aram Sinnreich, Patricia Aufderheide, and Rob Gehl. 2023. "Can This Platform Survive? Governance Challenges for the Fediverse." *SSRN Electronic Journal* (October 10, 2023). https://doi.org/10.2139/ssrn.4598303.

Suzor, Nicolas, Bryony Seignior, and Jennifer Singleton. 2017. "Non-Consensual Porn and the Responsibilities of Online Intermediaries." *Melbourne University Law Review* 40, no. 3 (January 1, 2017): 1057–97. https://law.unimelb.edu.au/__data/assets/pdf_file/0007/2329396/Suzor-403-Advance.pdf.

"The Bad Space." n.d. Accessed January 19, 2024. https://tweaking.thebad.space/about.

The Bluesky Team. 2023. "Towards Federation and an Open Network," November 15, 2023. https://blueskyweb.xyz/blog/11-15-2023-toward-federation.

thefuturebird. 2023. "Moderation: Sharing the load, Sharing tickets, feedback and more," January 7, 2023. https://github.com/mastodon/mastodon/discussions/22979.

Thiel, David, and Renée DiResta. 2023. *Child Safety on Federated Social Media.* Technical report. Stanford Internet Observatory, July 24, 2023. https://doi.org/10.25740/vb515nd6874.

Thomas, Kurt, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, et al. 2021. "SoK: Hate, Harassment, and the Changing Landscape of Online Abuse." *2021 IEEE Symposium on Security and Privacy (SP),* 247–67. https://doi.org/10.1109/SP40001.2021.00028.

Tworek, Heidi, and Alicia Wanless. 2022. "Time for Transparency from Digital Platforms, But What Does That Really Mean?" *Lawfare* (January 20, 2022). https://www.lawfaremedia.org/article/time-transparency-digital-platforms-what-does-really-mean.

Vilk, Viktorya, and Kat Lo. 2023. *Shouting into the Void.* Research report. PEN America, June 29, 2023. https://pen.org/report/shouting-into-the-void/.

Wanless, Alicia, and Michael Berk. 2020. "The Audience is the Amplifier: Participatory Propaganda." In *The SAGE Handbook of Propaganda,* edited by Paul Baines, Nicholas O'Shaughnessy, and Nancy Snow, 85–104. January 21, 2020. https://doi.org/10.4135/9781526477170.

Wardle, Claire, and Hossein Derahkshan. 2017. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking.* Research report. Council of Europe report DGI(2017)09, September 27, 2017. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Werdmuller, Ben. 2022. “Moderation on Mastodon: there’s a lot of work to do.” *werd.io* (November 21, 2022). https://werd.io/2022/moderation-on-mastodon-theres-a-lot-of-work-to-do.

X. 2022. “Sharing our latest transparency update, marking decade long commitment,” July 28, 2022. https://blog.twitter.com/en_us/topics/company/2022/ttr-20.

X Engineering. 2014. “Fighting spam with Botmaker,” August 20, 2014. https://blog.twitter.com/engineering/en_us/a/2014/fighting-spam-with-botmaker.

xg15. 2022. “I think the real danger for the Mastodon/greater Fediverse communtiy is...,” November 10, 2022. Accessed January 23, 2024. https://news.ycombinator.com/item?id=33545541.

Zuckerman, Ethan, and Chand Rajendra-Nicolucci. 2020. “What if Social Media Worked More Like Email?” *Knight First Amendment Institute at Columbia University* (November 3, 2020). https://knightcolumbia.org/blog/what-if-social-media-worked-more-like-email.

Authors

Yoel Roth (yoel@yoyoel.com) is a Knight Visiting Scholar at the University of Pennsylvania, a Technology Policy Fellow at the University of California, Berkeley, and a Nonresident Scholar at the Carnegie Endowment for International Peace. He was previously the head of trust and safety at Twitter, and received his PhD from the Annenberg School for Communication at the University of Pennsylvania.

Samantha Lai (samantha.lai@ceip.org) is a senior research analyst with the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace. Prior to joining Carnegie, she was a research analyst at the Brookings Institution and the Foreign Policy Research Institute.

Acknowledgements

The authors are grateful to Patrick Conlon, Renée DiResta, Camille François, Jeff Jarvis, Jaz-Michael King, Hilary Ross, and members of the Atlantic Council Task Force for a Trustworthy Future Web for their feedback and perspectives on earlier drafts of this article, and to Iryna Adam, Jon Bateman, Emma Landi, Alicia Wanless, Gavin Wilde, Kamya Yadav, and Joshua Sullivan for their support in facilitating interviews and analysis. The authors would also like to thank Agustina Del Campo, Andy Piper, Bryan Newbold, Darius Kazemi, David Thiel, Denise Paolucci, Derek Slater, Emelia Smith, Jaz-Michael King, Kate Klonick, Leigh Honeywell, Olga Belogolova, Quintessence Anx, Ross Schulman, Ted Han, Victoire Rio, and the other participants in our interview series for their time and expertise.

Data availability statement

Not applicable.

Funding statement

This work is supported by the William and Flora Hewlett Foundation, the John S. and James L. Knight Foundation, and the Carnegie Endowment for International Peace.

The Carnegie Endowment for International Peace's Partnership for Countering Influence Operations is grateful for funding provided from the Government of Canada, the William and Flora Hewlett Foundation, Craig Newmark Philanthropies, the John S. and James L. Knight Foundation, Microsoft, Meta, Google, Twitter, and WhatsApp. PCIO is wholly and solely responsible for the contents of its products, written or otherwise, and does not allow donors prior approval of drafts, influence on selection of project participants, or any influence over the findings and recommendations of work they may support.

Yoel Roth, the first author, was the Head of Trust and Safety at Twitter until November 2022. Since leaving Twitter, Roth has served as an independent consultant to a number of technology and social media companies regarding trust and safety issues, including one of the platforms evaluated as part of this article. This platform did not play any role in the study's design or data analysis, nor did it have the opportunity to approve the article before submission or publication. He serves on the advisory board of Independent Federated Trust and Safety (IFTAS), whose efforts are referenced in this article.

Ethical standards

Not applicable.

Keywords

Federated services; trust and safety; platform governance; content moderation.

Appendices

Appendix A: Questions for Semi-Structured Interviews

Group Interview 1: Community Risks and Responses

- What are the challenges that server administrators and moderators face on federated services?
- How could community moderators conduct actor-level analysis on campaigns like Spamouflage Dragon on federated services (Nimmo, Hubert, and Cheng 2021; Butler and Taege 2023)?
- How can community members of federated services address high-volume, low-sophistication political manipulation campaigns?
- It is seldom apparent from content alone that what you are looking at is part of a manipulative campaign. Take for example “Crystal Johnson,” an IRA persona purporting to be an African-American woman during the 2016 U.S. presidential elections (Giridharadas 2022). How could that be addressed on federated services?

Group Interview 2: Technological Risks and Responses

- How can we improve the detection and removal of CSAM on federated services?
- How do threat analysts conduct their work on centralized platforms? How would this be different on federated platforms?
- Can open-source tooling be a viable way to expand moderation support for federated services?
- What technical tools are required to support longitudinal analysis of inauthentic behavior on federated services?
- What kinds of antispam measures and tools do we need in federated services?
- What media hashing and matching functionalities already exist? What tools and infrastructure could be set up to help these functionalities evolve with the creation of new content?

Group Interview 3: Institutional Needs and Responses

- What are the needs of moderators on federated services?
- How are larger, existing institutions, such as Mozilla, engaged with federation? How do they contemplate institutional responses to the moderation challenges identified?
- What are existing models for international research collaboration, and how could these models apply to studying and moderating federated services?
- How might transparency reporting work for federated services? What kinds of institutional arrangements might support robust transparency reporting?
- How might institutional arrangements help support the development of appropriate safeguards to ensure that instance admins, or their designees, engage only in appropriate uses of sensitive user logs?

- What kinds of training and assistance can be provided to help moderators and users recognize inauthentic behavior? How might new institutional arrangements most productively be structured to develop these capacities?

Appendix B: Definitions

Publicly available community standards A public/user-facing document that defines rules or guidelines for behavior and content that is (and is not) acceptable on a platform. These documents may overlap with terms of service or other legal agreements, but must specifically concern user-generated content and conduct standards.

Publicly available specific policy definitions and enforcement criteria For policy areas prohibited under community standards, a public/user-facing document that provides specific guidelines about how violative content or conduct is defined by the platform, and how policies are enforced by the platform (e.g., content takedown, account removal).

Platform manipulation/behavioral policies and enforcement criteria Platform manipulation includes coordinated inauthentic activity, commercially motivated spam, or similar behaviors intended to artificially amplify or suppress information. These policies should include guidelines on how platform manipulation is defined, and how policies are enforced (e.g., content takedown, account removal).

User reporting capabilities for policy violations Users have the option to report posts/accounts for violating platform policies. Reporting options should include all or nearly all of the policies in the platform's community standards, or include a catch-all/other option.

Permanent account bans The ability of platform staff/moderators to permanently ban an account from accessing or posting on the platform. This can include variants on permanent bans that include read-only access.

Temporary account bans The ability of platform staff/moderators to temporarily prevent an account from accessing or posting on a platform for a specific period of time.

Ban evasion detection The ability of platform staff/moderators or automated platform systems to detect when a previously banned account has returned to the platform.

Post/content deletion The ability of platform staff/moderators to delete a post/content that violates community standards.

Account visibility restriction The ability of platform staff/moderators to restrict the appearance of an account to other users, in public/shared product surfaces (such as search), or in account recommendations.

Post/content visibility restriction The ability of platform staff/moderators to restrict the appearance of a post to other users, in public/shared product surfaces (such as search), or in content recommendations.

Demonetization The ability of platform staff/moderators to block a user from accessing post or account monetization features (such as ad revenue sharing), or block payouts to users deemed ineligible for monetization.

Automated enforcement tools (heuristics, ML) The use of automated enforcement techniques (including machine learning models, AI moderation systems, heuristics, or other automated processes) to detect and/or enforce against accounts/content that violates platform policy.

URL blocking The ability of platform staff/moderators to block users from accessing a URL through the platform, or prevent a URL from being shared in posts created on the platform.

Media hashing and matching The ability of platform staff/moderators to automatically detect if an uploaded picture matches with others in a repository of previously identified violative content. This is often used to detect Child Sexual Assault Material (CSAM).

User-facing moderation controls (block, mute, etc.) The ability of a user to prevent another user from interacting with them or hide all content/actions created by or associated with a target user. Specific affordances of block/mute features may vary by platform.

User identity verification (ID checks, etc.) The ability of platform staff/moderators to require users to provide information/evidence of offline/off-platform identity, such as by uploading a government-issued identification document (passport, driver's license, etc.) or "selfie" liveness verification.

Antispam challenges (reCAPTCHA, phone verification) The ability of platform staff/moderators to require accounts to complete a proof-of-humanness task, such as a captcha challenge or phone verification. Antispam challenges need not definitively/conclusively establish account identity.

Defederation/instance blocking The ability of servers' moderators to block the display of content created by accounts hosted on another server, or prevent another server from accessing content created by local users.

Published transparency report A public/user-facing document or website that provides information, including at least some specific quantitative data, about a company's trust and safety/content moderation actions.

Terms of service enforcement data A public/user-facing document or website that provides information, including at least some specific quantitative data, about a company's actions to enforce its terms of service (e.g., number of accounts taken down for violating a given policy).

Platform manipulation data A public/user-facing document or website that provides information, including at least some specific quantitative data, about a company's actions to enforce its terms of service regarding platform manipulation and other behavioral policies (e.g., account removals, attribution).

Legal information requests data A public/user-facing document or website that provides information, including at least some specific quantitative data, about a company's actions to respond to and/or comply with requests for user information submitted by external parties (including governments, civil litigants, etc.).

Legal removal demands data A public/user-facing document or website that provides information, including at least some specific quantitative data, about a company's actions to respond to and/or comply with requests submitted by external parties (including governments, civil litigants, etc.) to remove, restrict, or withhold user-generated content.

Country or jurisdictional breakdowns of data A published transparency report (as defined above) that provides country-specific (or other jurisdiction-level categories, as appropriate) data about platform actions.

Appendix C: Sources

Publicly available community standards

Facebook

- https://web.archive.org/web/20240120071152/https://transparency.fb.com/policies/community-standards/

Instagram

- https://web.archive.org/web/20240105051441/https://help.instagram.com/477434105621119

Horizon Worlds

- https://drive.google.com/file/d/1Hv45YC6UaJkjp_pgR1sb97p9J0BXmoAU/view

X

- https://web.archive.org/web/20240119002854/https://help.twitter.com/en/rules-and-policies/x-rules

Reddit

- https://web.archive.org/web/20240123004839/https://www.redditinc.com/policies/content-policy

YouTube

- https://web.archive.org/web/20240122002558/https://www.youtube.com/howyoutubeworks/policies/community-guidelines/

Threads

- https://web.archive.org/web/20240122151228/https://about.fb.com/news/2023/07/introducing-threads-new-app-text-sharing/

Bluesky

- https://web.archive.org/web/20240101050917/https://blueskyweb.xyz/support/community-guidelines

Mastodon

- https://web.archive.org/web/20240122143356/https://joinmastodon.org/covenant

Pixelfed

- https://web.archive.org/web/20231127155715/https://pixelfed.social/site/kb/community-guidelines

diaspora

- https://web.archive.org/web/20240123154636/https://discourse.diasporafoundation.org/faq
- ttps://web.archive.org/web/20231003043409/https://diasporafoundation.org/community_guidelines

PeerTube

- https://web.archive.org/web/20231126220921/https://docs.joinpeertube.org/contribute/code-of-conduct

Publicly available specific policy definitions and enforcement criteria

Facebook/Instagram

- https://web.archive.org/web/20231206173544/https://transparency.fb.com/enforcement/

X

- https://web.archive.org/web/20240118060043/https://help.twitter.com/en/rules-and-policies
- https://web.archive.org/web/20240123090005/https://help.twitter.com/en/rules-and-policies/enforcement-options

YouTube

- https://web.archive.org/web/20240121114618/https://support.google.com/youtube/answer/2802032

- https://web.archive.org/web/20240122140312/https://support.google.com/youtube/answer/2802168

Threads

- https://drive.google.com/file/d/1wPsQ3JYa3mJAx2G-lat_GWc8j08HuNCV/view
- While the Threads terms of use point back to the “Instagram Terms of Use and Instagram Community Guidelines,” it remains unclear how broadly Instagram’s guidelines and enforcement procedures are applied to content created and shared on Threads.

Bluesky

- https://web.archive.org/web/20240119082119/https://blueskyweb.xyz/blog/4-13-2023-moderation

Pixelfed

- https://web.archive.org/web/20231127155715/https://pixelfed.social/site/kb/community-guidelines

PeerTube

- https://web.archive.org/web/20231126220921/https://docs.joinpeertube.org/contribute/code-of-conduct

Platform manipulation/behavioral policies and enforcement criteria

Facebook

- https://web.archive.org/web/20230909100702/https://transparency.fb.com/policies/community-standards/inauthentic-behavior/

Instagram

- https://web.archive.org/web/20240118024543/https://help.instagram.com/477434105621119/

X

- https://web.archive.org/web/20240114085509/https://help.twitter.com/en/rules-and-policies/platform-manipulation

Reddit

- https://web.archive.org/web/20240123155955/https://www.redditinc.com/policies/content-policy

Youtube

- https://web.archive.org/web/20231128052129/https://support.google.com/youtube/answer/3399767

Threads

- https://drive.google.com/file/d/1wPsQ3JYa3mJAx2G-lat_GWc8j08HuNCV/view
- While the Threads terms of use point back to the “Instagram Terms of Use and Instagram Community Guidelines,” it remains unclear how broadly Instagram’s guidelines and enforcement procedures are applied to content created and shared on Threads.

User reporting capabilities for policy violations

Facebook

- https://drive.google.com/file/d/1ajXWru8rYqy6I84ZWD7pKgphJSaUvrtJ/view

Instagram

- https://drive.google.com/file/d/15aM4OrM8JpXvCBFAVltTbvzWb5dB-vuk/view

Horizon Worlds

- https://drive.google.com/file/d/1gtAF1M1TntCQtz8it4tgl02CcDGEC8TN/view

X

- https://web.archive.org/web/20240121164312/https://help.twitter.com/en/rules-and-policies/x-report-violation

Reddit

- https://web.archive.org/web/20230712054802/https://support.reddithelp.com/hc/en-us/articles/360058309512-How-do-I-report-a-post-or-comment-

YouTube

- https://web.archive.org/web/20240110041400/https://support.google.com/youtube/answer/2802027

Threads

- https://web.archive.org/web/20231003043735/https://help.instagram.com/6602413966453273

Bluesky

- https://web.archive.org/web/20240101050917/https://blueskyweb.xyz/support/community-guidelines

Mastodon

- https://web.archive.org/web/20231208205033/https://docs.joinmastodon.org/entities/Report/

Pixelfed

- https://web.archive.org/web/20220924194946/https://pixey.org/site/kb/safety-tips
- https://drive.google.com/file/d/1BFAF-V5N8zwOI5n8TnTciXoZtpZLpS73/view

diaspora

- https://web.archive.org/web/20230926113321/https://blog.diasporafoundation.org/5-dealing-with-problem-content-in-a-distributed-system
- https://web.archive.org/web/20240123161315/https://discourse.diasporafoundation.org/t/confused-by-reported-post/1988/4

PeerTube

- https://web.archive.org/web/20231204132427/https://docs.joinpeertube.org/admin/moderation

Permanent account bans

Facebook

- https://web.archive.org/web/20230824002339/https://transparency.fb.com/enforcement/taking-action/restricting-accounts/

Instagram

- https://web.archive.org/web/20230314223717/https://help.instagram.com/366993040048856
- https://web.archive.org/web/20230926215719/https://inssist.com/knowledge-base/instagram-bans-blocks-and-limits

Horizon Worlds

- https://drive.google.com/file/d/1Hv45YC6UaJkjp_pgR1sb97p9J0BXmoAU/view
- https://drive.google.com/file/d/1lDg4pHULfwMprJeFjCmHHVmHXlDXTdS7/view
- https://drive.google.com/file/d/1rrzafWZJV_1BGHvrVYldCabnJwvWhwok/view

X

- https://web.archive.org/web/20240117064100/https://help.twitter.com/en/rules-and-policies/notices-on-x

Reddit

- https://web.archive.org/web/20230329051752/https://reddit.zendesk.com/hc/en-us/articles/360045734911-My-account-has-been-permanently-suspended
- https://drive.google.com/file/d/1DD5tmcVGbUnRrQG0JuYm6hPfPm5xH929/view

YouTube

- https://web.archive.org/web/20240110041442/https://support.google.com/youtube/answer/2802168

Threads

- https://drive.google.com/file/d/1wPsQ3JYa3mJAx2G-lat_GWc8j08HuNCV/view
- https://web.archive.org/web/20230314223717/https://help.instagram.com/366993040048856
- https://drive.google.com/file/d/1m0fpub2hvRL7XtgPbZnZI5xVwohSw804/view

Bluesky

- https://web.archive.org/web/20240112103701/https://blueskyweb.xyz/support/tos#general-prohibitions-enforcement

Mastodon

- https://web.archive.org/web/20240119041327/https://docs.joinmastodon.org/admin/moderation/

Pixelfed

- https://web.archive.org/web/20231127155715/https://pixelfed.social/site/kb/community-guidelines

diaspora

- https://web.archive.org/web/20230705161111/https://discourse.diasporafoundation.org/tos
- https://web.archive.org/web/20230926113321/https://blog.diasporafoundation.org/5-dealing-with-problem-content-in-a-distributed-system
- https://web.archive.org/web/20231003043409/https://diasporafoundation.org/community_guidelines
- Notably, while the Diaspora Terms of Service note that platforms can terminate any account at any time, the service's "social principles" page states that the platform does not allow for the banning of members. We interpret this to mean that the capacity and technical capability to ban users exists, but as a matter of practice is not used by moderators.

PeerTube

- https://web.archive.org/web/20231204132427/https://docs.joinpeertube.org/admin/moderation
- https://web.archive.org/web/20190523122901/https://github.com/Chocobozzz/PeerTube/issues/718

Temporary account bans/timeouts

Facebook

- https://web.archive.org/web/20230824002339/https://transparency.fb.com/enforcement/taking-action/restricting-accounts/

Instagram

- https://drive.google.com/file/d/1ZB9Wyvi7G-aBzIPB-nbgzFdYv_McMonx/view

Horizon Worlds

- https://drive.google.com/file/d/1rrzafWZJV_1BGHvrVYldCabnJwvWhwok/view
- https://web.archive.org/web/20230403152832/https://about.fb.com/news/2022/07/meta-accounts-and-horizon-profiles-for-vr/

X

- https://web.archive.org/web/20240117064100/https://help.twitter.com/en/rules-and-policies/notices-on-x

Reddit

- https://web.archive.org/web/20230330140023/https://reddit.zendesk.com/hc/en-us/articles/360045308832-My-account-has-been-temporarily-suspended

Youtube

- https://web.archive.org/web/20240122140312/https://support.google.com/youtube/answer/2802168

Threads

- https://drive.google.com/file/d/1wPsQ3JYa3mJAx2G-lat_GWc8j08HuNCV/view

Bluesky

- https://web.archive.org/web/20240122085813/https://blueskyweb.xyz/support/tos#general-prohibitions-enforcement

Mastodon

- https://web.archive.org/web/20240119041327/https://docs.joinmastodon.org/admin/moderation/
- https://web.archive.org/web/20240123162557/https://github.com/mastodon/mastodon/issues/11328

Pixelfed

- https://web.archive.org/web/20231127155715/https://pixelfed.social/site/kb/community-guidelines
- https://drive.google.com/file/d/11gftAvF1aII03HnCBRWG9RoqqQ8-QuVw/view

PeerTube

- https://web.archive.org/web/20230325051038/https://docs.joinpeertube.org/contribute/code-of-conduct
- https://web.archive.org/web/20240123162534/https://github.com/Chocobozzz/PeerTube/issues/5101

Ban evasion detection

Facebook

- https://web.archive.org/web/20240122102749/https://transparency.fb.com/policies/community-standards/account-integrity-and-authentic-identity/
- https://web.archive.org/web/20240112205100/https://transparency.fb.com/policies/ad-standards/business-assets/evading-enforcement

Instagram

- https://web.archive.org/web/20240122102749/https://transparency.fb.com/policies/community-standards/account-integrity-and-authentic-identity/
- https://web.archive.org/web/20230908231451/https://incogniton.com/instagram-ip-ban/

Horizon Worlds

- Users need a Meta account to set up a Horizon Worlds account. It therefore stands to reason that Facebook-specific or Meta-wide measures for detecting ban evasion would apply to Meta accounts used for Horizon Worlds.

X

- https://web.archive.org/web/20231123044830/https://help.twitter.com/en/rules-and-policies/ban-evasion

Reddit

- https://web.archive.org/web/20230610134731/https://www.reddit.com/r/modnews/comments/wrnnvb/piloting_a_new_ban_evasion_tool/

YouTube

- https://web.archive.org/web/20231205235132/https://support.google.com/youtube/answer/2802032

Threads

- The Threads terms of use point back to the “Instagram Terms of Use and Instagram Community Guidelines.” Per discussions with Meta representatives, this enforcement procedure is applied to content created and shared on Threads.

Mastodon

- https://drive.google.com/file/d/1Cy0NX3wCxLMQs_oMlerzuXqQw9RkNjzm/view

Pixelfed

- https://web.archive.org/web/20240116235855/https://pixelfed.social/site/privacy

PeerTube

- https://web.archive.org/web/20240123162941/https://framagit.org/rigelk/peertube-plugin-glavlit/-/issues/1

Post/content deletion

Facebook

- https://web.archive.org/web/20230617163151/https://transparency.fb.com/enforcement/taking-action/taking-down-violating-content/

Instagram

- https://drive.google.com/file/d/1fhW3MZosnrBz20rW5Pw67h8CRCZ_BNBG/view

X

- https://web.archive.org/web/20231228091844/https://help.twitter.com/en/rules-and-policies/enforcement-options

Reddit

- https://web.archive.org/web/20231123111135/https://www.redditinc.com/policies/user-agreement-september-12-2021

YouTube

- https://web.archive.org/web/20231207105925/https://support.google.com/youtube/answer/6395024
- https://web.archive.org/web/20240121213529/https://www.youtube.com/howyoutubeworks/our-commitments/managing-harmful-content/#reduce

Threads

- https://drive.google.com/file/d/1T8Qq5wh9p46JAwSsp56B_h_ERYkfz0en/view
- https://drive.google.com/file/d/1fhW3MZosnrBz20rW5Pw67h8CRCZ_BNBG/view?usp=drive_link

Bluesky

- https://web.archive.org/web/20240101050917/https://blueskyweb.xyz/support/community-guidelines

Mastodon

- https://drive.google.com/file/d/1QzzRb9nNASJAD2CmwP2kHDczZtknbfw4/view

Pixelfed

- https://web.archive.org/web/20231127155715/https://pixelfed.social/site/kb/community-guidelines

diaspora

- https://web.archive.org/web/20240123154636/https://discourse.diasporafoundation.org/faq

PeerTube

- https://web.archive.org/web/20231204132427/https://docs.joinpeertube.org/admin/moderation

Account visibility restriction

Facebook

- https://web.archive.org/web/20240113181655/https://transparency.fb.com/fi-fi/enforcement/taking-action/restricting-accounts/

Instagram

- https://web.archive.org/web/20230705005006/https://help.instagram.com/539126347315373
- https://web.archive.org/web/20230531121902/https://help.instagram.com/313829416281232?helpref=faq_content

X

- https://web.archive.org/web/20240117064100/https://help.twitter.com/en/rules-and-policies/notices-on-x

Reddit

- https://web.archive.org/web/20240123155955/https://www.redditinc.com/policies/content-policy https://web.archive.org/web/20220923102759/https://www.reddit.com/r/NewToReddit/comments/xh73wb/i_created_a_new_account_but_my_posts_and_comments/

YouTube

- https://web.archive.org/web/20240121114618/https://support.google.com/youtube/answer/2802032

Threads

- https://web.archive.org/web/20230705005006/https://help.instagram.com/539126347315373
- The Threads terms of use point back to the "Instagram Terms of Use and Instagram Community Guidelines." While it includes guidelines on best practices for accounts, it remains unclear how these capabilities are applied to content created and shared on Threads.

Bluesky

- https://drive.google.com/file/d/1pgj-dOL7Ir1Q6tONHP6nJBmIOuhS7Z-6/view

Mastodon

- https://web.archive.org/web/20240119041327/https://docs.joinmastodon.org/admin/moderation/#limit-user

Pixelfed

- https://web.archive.org/web/20231127155715/https://pixelfed.social/site/kb/community-guidelines

PeerTube

- https://web.archive.org/web/20231126220930/https://docs.joinpeertube.org/use/mute

Post/content visibility restriction

Facebook

- https://web.archive.org/web/20240113181655/https://transparency.fb.com/fi-fi/enforcement/taking-action/restricting-accounts/

Instagram/Threads

- https://drive.google.com/file/d/1TErmReFlWXP-xbxmW-gMJ6_FyWCKC_VJ/view

X

- https://web.archive.org/web/20231218011944/https://help.twitter.com/en/rules-and-policies/notices-on-x

Reddit

- https://web.archive.org/web/20240123155955/https://www.redditinc.com/policies/content-policy
- https://web.archive.org/web/20231202013824/https://support.reddithelp.com/hc/en-us/articles/360043069012

YouTube

- https://web.archive.org/web/20240120042613/https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/

Threads

- https://drive.google.com/file/d/1EjWYzZZfq0K2lTxDwBuELykv-Hlvtib_/view
- https://drive.google.com/file/d/1H6nRx8kPcumq3PGTA8epwQRQJ7nEguPY/view

Bluesky

- https://web.archive.org/web/20240115204023/https://blueskyweb.xyz/faq

Mastodon

- https://web.archive.org/web/20240119041327/https://docs.joinmastodon.org/admin/moderation/#limit-user
- https://drive.google.com/file/d/1QzzRb9nNASJAD2CmwP2kHDczZtknbfw4/view

Pixelfed

- https://drive.google.com/file/d/1bC56z3wC1pOmn3CAqpwGtbH0tJZiZmdz/view

PeerTube

- https://web.archive.org/web/20231126220930/https://docs.joinpeertube.org/use/mute

Demonetization

Facebook

- https://drive.google.com/file/d/1bOoaJwYrXjJkclSXcg8K7cV_St5AxbsC/view

Instagram

- https://drive.google.com/file/d/1sZGyDwHTuH6TCXs1-nIfDmYaoZbcz-a4/view

Horizon Worlds

- https://drive.google.com/file/d/1-3vc8uddUwlN9qmG7AXIGdaUDi0kM5q1/view

X

- https://web.archive.org/web/20240109211854/https://help.twitter.com/en/rules-and-policies/content-monetization-standards

YouTube

- https://web.archive.org/web/20240121210249/https://support.google.com/youtube/answer/1311392

Automated enforcement tools (heuristics, ML)

Facebook

- https://web.archive.org/web/20231006154154/https://about.fb.com/news/2020/08/how-we-review-content/

Instagram

- https://drive.google.com/file/d/1PoanMxsRnT-rXtJ8K4cRYGpGhPZB-N7M/view

Horizon Worlds

- https://drive.google.com/file/d/1e7w6rNL9RU4p7rtPrwNQDXoiDqZdI7hE/view

X

- https://drive.google.com/file/d/1qs29REVGUdhyefUtdmO8OhF5pkAsBDQD/view
- https://drive.google.com/file/d/1KoVubdtucmUfPYRTC8nCb1uOlL5Yo1nF/view

Reddit

- https://web.archive.org/web/20240117022057/https://www.reddit.com/wiki/automoderator/

YouTube

- https://web.archive.org/web/20240120042614/https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/

Threads

- https://drive.google.com/file/d/1PoanMxsRnT-rXtJ8K4cRYGpGhPZB-N7M/view
- The Threads terms of use point back to the “Instagram Terms of Use and Instagram Community Guidelines.” Per discussions with Meta representatives, this enforcement procedure is applied to content created and shared on Threads.

Bluesky

- https://web.archive.org/web/20240115204023/https://blueskyweb.xyz/faq

Pixelfed

- https://web.archive.org/web/20240105020534/https://docs.pixelfed.org/technical-documentation/config/#captcha

diaspora

- https://web.archive.org/web/20240123205102/https://discourse.diasporafoundation.org/t/discourse-sent-a-message-that-my-question-was-a-spam/2618/2

URL blocking

Facebook

- https://drive.google.com/file/d/1kAldyFaGoyDrl4u1EZ5Ek4a96byjVMIg/view

Instagram

- https://drive.google.com/file/d/1AphUBimvUCy1rzLhHSegkscB-c7zN3ZX/view
- https://drive.google.com/file/d/16d4UxnAmVi0vIxSQIa7V8HUmrwx9UYsT/view
- The Threads terms of use point back to the "Instagram Terms of Use and Instagram Community Guidelines." Per discussions with Meta representatives, this enforcement procedure is applied to content created and shared on Threads.

X

- https://web.archive.org/web/20231223103946/https://help.twitter.com/en/safety-and-security/phishing-spam-and-malware-links

Reddit

- https://web.archive.org/web/20240117022057/https://www.reddit.com/wiki/automoderator/

YouTube

- https://web.archive.org/web/20240107215425/https://support.google.com/youtube/answer/9054257

Media hashing and matching

Facebook and Instagram

- https://web.archive.org/web/20240111021643/https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/
- https://web.archive.org/web/20240117020701/https://about.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/
- https://web.archive.org/web/20240123205140/https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images/

X

- Twitter/X have not publicly shared specific details of their systems, but are founding members of GIFCT, and are known to use GIFCT-derived hashes. See:https://gifct.org/hsdb/

Reddit

- https://web.archive.org/web/20230420022803/https://reddit.zendesk.com/hc/en-us/articles/10654543840276-How-does-Reddit-fight-Child-Sexual-Exploitation-

YouTube

- https://web.archive.org/web/20231204093719/https://transparencyreport.google.com/

Threads

- The Threads terms of use point back to the "Instagram Terms of Use and Instagram Community Guidelines." Per discussions with Meta representatives, this enforcement procedure is applied to content created and shared on Threads.

Bluesky

- https://drive.google.com/file/d/1zkV669SqgssAMDy9sr7g6ybC0ZsfWqVN/view
- https://web.archive.org/web/20240119085526/https://blueskyweb.xyz/blog/01-16-2024-moderation-2023
- Per discussions with Bluesky developers, Bluesky has operated both ML-based classification of visual images, and perceptual hash-based matching of visual images against CSAM indices, for all content since September 2023.

User-facing moderation controls (block, mute, etc.)

Facebook

- https://drive.google.com/file/d/1cRvrH-pG4leQHl8rAN_5s5bDFyTGw3nc/view

Instagram

- https://drive.google.com/file/d/1y76im7s_a8W3fjiEZN7DY1XC5mqP4Way/view

Horizon Worlds

- https://drive.google.com/file/d/1eVKApofZcIrgEZ_kIw3Dl7-mdzJwA-Ag/view

X

- https://web.archive.org/web/20231217122427/https://help.twitter.com/en/using-x/x-mute
- https://web.archive.org/web/20240108235132/https://help.twitter.com/en/using-x/blocking-and-unblocking-accounts

Reddit

- https://drive.google.com/file/d/1YD5_tdG7yt-7njI-4W6ZyFcZ-eHZnM_E/view

YouTube

- https://drive.google.com/file/d/1lfXzom0VNHuwRFJJvN4tMq-czyaP7S3c/view
- https://web.archive.org/web/20231230005124/https://support.google.com/youtube/answer/9482361

Threads

- https://web.archive.org/web/20231003044829/https://help.instagram.com/616605623708734/
- https://web.archive.org/web/20231223075217/https://help.instagram.com/179980294969821

Bluesky

- https://web.archive.org/web/20240114003515/https://blueskyweb.xyz/blog/5-19-2023-user-faq

Mastodon

- https://web.archive.org/web/20240114143301/https://docs.joinmastodon.org/user/moderating/

Pixelfed

- https://web.archive.org/web/20240105020534/https://docs.pixelfed.org/technical-documentation/config/#cs_blocked_actor

diaspora

- https://web.archive.org/web/20240123000444/https://wiki.diasporafoundation.org/FAQ_for_users

PeerTube

- https://web.archive.org/web/20231204132427/https://docs.joinpeertube.org/admin/moderation
- https://web.archive.org/web/20231208205023/https://docs.joinpeertube.org/admin/managing-users

User identity verification (ID checks, etc.)

Facebook

- https://drive.google.com/file/d/1Ub_gZ-PnQFkvrYssKEptkpobtqYZloXC/view

Instagram

- https://web.archive.org/web/20231004212325/https://help.instagram.com/271237319690904/
- https://drive.google.com/file/d/1AyezuB3dvRl9TZlIqNrWJk1_CiHaEkHQ/view

Horizon Worlds

- https://drive.google.com/file/d/12-a3VZwOj462qAY8Z4wqVggnfrz1ifG5/view

X

- https://web.archive.org/web/20230920004859/https://www.theverge.com/2023/9/15/23874854/x-twitter-verification-government-id-paid-account-benefits

Reddit

- https://drive.google.com/file/d/1K6ANMhijZSpzVPxUhR0UduPHzqZI1uuM/view
- Reddit has ID verification processes for advertisers only.

Youtube

- https://web.archive.org/web/20230403095610/https://support.google.com/youtube/answer/7644078

Bluesky

- https://web.archive.org/web/20231223143533/https://blueskyweb.org/blog/press-faq
- Bluesky does not have identity verification capabilities, but it does use domain verification as an alternative identity mechanism.

Mastodon

- https://web.archive.org/web/20240118111242/https://joinmastodon.org/verification
- Mastodon does not have identity verification capabilities, but it does use domain verification as an alternative identity mechanism.

Pixelfed

- https://drive.google.com/file/d/1IUIxYuUXcku_YWA3c0nVbgOgbxzF2Eus/view
- Pixelfed does not have identity verification capabilities, but it does use domain verification as an alternative identity mechanism.

Antispam challenges (reCAPTCHA, phone verification)

Facebook

- https://drive.google.com/file/d/18dFjVywyf_7OzIrSVPHOE6Oy979uJIgx/view

Instagram

- https://web.archive.org/web/20240123033829/https://help.instagram.com/477434105621119
- https://web.archive.org/web/20240109225400/https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram
- https://web.archive.org/web/20231223103919/https://help.instagram.com/165828726894770

X

- https://web.archive.org/web/20240114085509/https://help.twitter.com/en/rules-and-policies/platform-manipulation

Reddit

- https://drive.google.com/file/d/1LK_sOowhFl_Wjdf7xCsMp-DtWqS5y_2H/view

YouTube

- https://drive.google.com/file/d/1UwYPcWZwMxtVPes-kSK1FEw31JF9ZTB7/view

Threads

- https://web.archive.org/web/20230718055553/https://techcrunch.com/2023/07/17/the-spam-bots-have-now-found-threads-as-company-announces-its-own-rate-limits/

Mastodon

- https://web.archive.org/web/20240119041327/https://docs.joinmastodon.org/admin/moderation/#spam-fighting-measures

Pixelfed

- https://web.archive.org/web/20240105020534/https://docs.pixelfed.org/technical-documentation/config/#captcha
- https://web.archive.org/web/20230507171227/https://mastodon.social/@dansup/109908200001367488

diaspora

- https://web.archive.org/web/20240123213134/https://discourse.diasporafoundation.org/t/captcha-image-a-broken-link/1690
- https://web.archive.org/web/20240123213142/https://discourse.diasporafoundation.org/t/spam-control-for-podmins-could-someone-summarize-please/2748
- https://web.archive.org/web/20220701075134/https://discourse.diasporafoundation.org/t/add-stopforumspam-integration/2038
- https://web.archive.org/web/20210729031955/https://discourse.diasporafoundation.org/t/better-abilities-for-podmins-for-spam-analytics-and-controls/3896
- https://web.archive.org/web/20240123000451/https://wiki.diasporafoundation.org/FAQ_for_pod_maintainers

PeerTube

- https://web.archive.org/web/20231204132427/https://docs.joinpeertube.org/admin/moderation
- https://drive.google.com/file/d/1JKT7R3pYKd4qjqAK6PKKQ8_0ctHA-i7L/view

Defederation/instance blocking

Mastodon

- https://web.archive.org/web/20240123183205/https://docs.joinmastodon.org/user/moderating/#block-domain
- https://web.archive.org/web/20231208205029/https://docs.joinmastodon.org/methods/admin/domain_allows/

Pixelfed

- https://drive.google.com/file/d/1uVJ9ystAj8oJCYnxq6ulHTAyWfnkXrCE/view

PeerTube

- https://web.archive.org/web/20231204132427/https://docs.joinpeertube.org/admin/moderation

Published transparency report

Facebook/Instagram

- https://web.archive.org/web/20231222165758/https://transparency.fb.com/reports/

X

- https://web.archive.org/web/20240121233141/https://transparency.twitter.com/en/reports.html

Reddit

- https://web.archive.org/web/20231206054728/https://www.redditinc.com/policies/mid-year-transparency-report-2022

YouTube

- https://drive.google.com/file/d/1peWPf8l-dVKUvmaAkNKPrvBXGuzgxYxy/view

Threads

- As Threads was launched in July 2023, the company has not yet released a transparency report.

Bluesky

- https://web.archive.org/web/20240117113948/https://blueskyweb.xyz/blog/01-16-2024-moderation-2023

Mastodon

- https://web.archive.org/web/20240121225234/https://joinmastodon.org/reports/Mastodon%20Annual%20Report%202022.pdf

Terms of service enforcement data

Facebook/Instagram

- https://drive.google.com/file/d/17J4hkYdKt8D9t-aqoB2_9geh-Ofe041k/view

X

- https://web.archive.org/web/20240113231129/https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec
- https://drive.google.com/file/d/10gA1lUeOFew9rTTqnvwV6JmYHGOoe8CG/view

Reddit

- https://web.archive.org/web/20231206054728/https://www.redditinc.com/policies/mid-year-transparency-report-2022

Youtube

- https://drive.google.com/file/d/1sqauqDqX9bNM86KvwEbtIb3RWi-c00JT/view

Bluesky

- https://web.archive.org/web/20240117113948/https://blueskyweb.xyz/blog/01-16-2024-moderation-2023

Mastodon

- https://web.archive.org/web/20240121225234/https://joinmastodon.org/reports/Mastodon%20Annual%20Report%202022.pdf

Platform manipulation data

Facebook/Instagram

- https://web.archive.org/web/20240122120058/https://about.fb.com/news/tag/coordinated-inauthentic-behavior/

X

- https://web.archive.org/web/20231210010640/https://transparency.twitter.com/en/reports/platform-manipulation.html#2021-jul-dec
- https://drive.google.com/file/d/1zilgmN5XrUca4jNOXLK-VVHnXRZoJKeP/view

Reddit

- https://web.archive.org/web/20240116084824/https://www.redditinc.com/policies/transparency-report-2021-2/
- https://web.archive.org/web/20230621173713/https://www.reddit.com/r/redditsecurity/comments/rgikn1/q3_safety_security_report/

Legal information requests data

Facebook/Instagram

- https://drive.google.com/file/d/1mRj8Nq8jFbWdXUSwWSqMbthrVtCpUC1y/view

X

- https://web.archive.org/web/20240117151931/https://transparency.twitter.com/en/reports/information-requests.html#2021-jul-dec

- https://drive.google.com/file/d/10gA1lUeOFew9rTTqnvwV6JmYHGOoe8CG/view

Reddit

- https://web.archive.org/web/20240116084824/https://www.redditinc.com/policies/transparency-report-2021-2/

YouTube

- https://storage.googleapis.com/transparencyreport/google-user-data-requests.zip
- https://web.archive.org/web/20240113230724/https://transparencyreport.google.com/

Legal removal demands data

Facebook/Instagram

- https://drive.google.com/file/d/1la5LqFlMA0a6-YKzdQb8dFRTasYWmftw/view

X

- https://drive.google.com/file/d/11vpliILLBQ0QJ1mWmWpy3qdY0V3XfF0i/view
- https://drive.google.com/file/d/10gA1lUeOFew9rTTqnvwV6JmYHGOoe8CG/view

Reddit

- https://drive.google.com/file/d/1P4VjUvZJzyzrLIAZN_tC9w5tJnwmA9NX/view

YouTube

- https://drive.google.com/file/d/1aOPdvl-BoiLkfAjcgB5ettCtICSR3-oG/view
- https://storage.googleapis.com/transparencyreport/google-government-removals.zip

Country or jurisdictional breakdowns of data

Facebook/Instagram

- https://drive.google.com/file/d/1oVLHNubAENI2z6tIQygxdHEh45ivZJwX/view?usp=drive_link

X

- https://web.archive.org/web/20240117151931/https://transparency.twitter.com/en/reports/information-requests.html#2021-jul-dec
- https://blog.twitter.com/en_us/topics/company/2023/an-update-on-twitter-transparency-reporting

Reddit

- https://web.archive.org/web/20240116084824/https://www.redditinc.com/policies/transparency-report-2021-2/

YouTube

- https://drive.google.com/file/d/1IjGubhJqe0iPeVZjPRktHch1PNquU5Jx/view