Spencer-Smith, Charlotte

Article

Labour pains: Content moderation challenges in Mastodon growth

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Spencer-Smith, Charlotte (2025) : Labour pains: Content moderation challenges in Mastodon growth, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 14, Iss. 1, pp. 1-21, https://doi.org/10.14763/2025.1.1831

This Version is available at:
https://hdl.handle.net/10419/315588

INTERNET POLICY REVIEW

Volume 14 **Issue** 1

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Labour pains: Content moderation challenges in Mastodon growth

Charlotte Spencer-Smith *University of Klagenfurt*
Tales Tomaz *University of Salzburg*

DOI: https://doi.org/10.14763/2025.1.1831

Published: 31 March 2025
Received: 17 March 2024 **Accepted:** 7 August 2024

Funding: The authors did not receive any funding for this research.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Spencer-Smith, C., & Tomaz, T. (2025). Labour pains: Content moderation challenges in Mastodon growth. *Internet Policy Review*, *14*(1). https://doi.org/10.14763/2025.1.1831

Keywords: Mastodon, Content moderation, Fediverse, Social media

Abstract: After Elon Musk took over Twitter in October 2022, the number of users on the alternative social media platform Mastodon rose dramatically. The sudden influx of new users posed several challenges to content moderation distinct from those in large commercial social media. This article investigates the challenges Mastodon communities have faced and how their admins and content moderators have managed them. Based on scholarly literature, the article contextualises Mastodon as an open source, federated alternative to corporate social media and explains how content moderation is expected to occur in this model, including possible challenges from sudden growth in user numbers. The article then empirically investigates challenges experienced by Mastodon instances post-Musk, based on eight interviews with admins and moderators of seven instances and a representative of Independent Federated Trust & Safety (IFTAS), a non-profit organisation that supports Mastodon content moderators. The research finds that challenges and the responses to them vary depending on the characteristics of the instance, such as size, thematic focus and geography, andinstances tend to adopt measures tailored to their communities. However, a tension between centralisation and decentralisation, including Global North-South differences, cuts across the network, which may be accentuated by further growth.

This paper is part of **Content moderation on digital platforms: beyond states and firms**, a special issue of *Internet Policy Review* guest-edited by Romain Badouard and Anne Bellon.

Introduction

In the two months after Elon Musk became the owner of Twitter – now rebranded as X –, in late October 2022, users became increasingly concerned about his management of the platform, particularly the quality of its content moderation (Farrell, 2022). As users sought alternatives, the number of monthly active users on the alternative social media platform Mastodon rose from 300,000 to 2.5 million (Peters, 2022). Mastodon has emerged as an open source alternative to corporate social media platforms that relies on the concept of federation: social networks should use open and interoperable protocols and software, operating independently but still allowing their users to reach people with accounts in other networks (Gehl, 2015; Zulli et al., 2020). This technical affordance has given rise to thousands of independent networks using Mastodon, as well as other interoperable applications. Each of them is called an *instance*. They have different sizes in terms of users, ranging from one-person instances to hundreds of thousands or even millions of registered users. Mastodon content moderation occurs mostly on the level of these instances, instead of being enforced by one central entity, as in corporate social media.

However, the recent significant growth in user numbers has posed challenges for the content moderation infrastructures of Mastodon instances. In this context, this article investigates the following research questions: 1) Which challenges emerged for content moderation in Mastodon communities after the takeover of Twitter by Elon Musk? 2) How have admins and content moderators responded to these challenges?

First, the article contextualises Mastodon within the free software movement and its quest for digital sovereignty. Secondly, the article explains how content moderation is conceived in this model. Then, the article discusses the challenges that sudden growth in user numbers presents to content moderation in social media. Finally, the empirical part investigates how these challenges affected Mastodon in its user growth after Twitter/X's takeover by Elon Musk, based on eight interviews with admins and moderators of seven instances and a representative of Independent Federated Trust & Safety (IFTAS), a non-profit organisation that supports Mastodon content moderators.

From corporate to federated social media

Social media platforms have become a key element of contemporary communication systems (Flensburg & Lai, 2020). After an early phase of optimism, they now are held responsible for several problems such as the spread of misinformation, disinformation, hate speech and surveillance (Miller & Vaccari, 2020). Whereas it is clear that malicious actors are abusing technologies, there is also a widespread understanding that structural features in current social media platforms provide incentives to harmful behaviour (Griffin, 2023; Rahman & Teachout, 2020). Hence, stakeholders such as entrepreneurs, scholars, policymakers, activists and civil society organisations, not to mention the platforms themselves, offer different explanations and solutions for these issues.

One proposal comes from the Free and Open Source Software (FOSS) movement, namely to create alternative social networks that can be locally managed but globally connected or 'federated'. The contributors to FOSS projects are often volunteers, who in lieu of financial remuneration, have other motivations to participate, such as pursuing personal interests or furthering a perceived public or community good (Butler et al,. 2007; Terry et al., 2010). With open source protocols and software, management of social media becomes distributed to smaller groups, who can choose how and when to interact with each other, resembling a federated system. This proposal has given birth to the so-called "Fediverse", an assemblage of federated social media developed over the last 15 years (Gehl, 2015; Rozenshtein, 2023; Zulli et al., 2020).

The ActivityPub protocol and Mastodon have become the most successful iterations of this process. ActivityPub was introduced in January 2018 and became the World Wide Web Consortium's (W3C) standard recommendation for social media. It offers a basic grammar for typical social media activities, such as posting and liking. Built largely by queer developers, its features are conceived to protect vulnerable communities, who are often harassed and abused under the free speech absolutism of commercial platforms (Klemens, 2023). Crucially, the ActivityPub protocol enables a social network operating on it to be interoperable with other social networks that use the same protocol. In practical terms, Mastodon, for example, is now interoperable with Instagram Threads, even though these networks are distinct from one another, as they both use ActivityPub (Pierce, 2024).

ActivityPub was quickly adopted by Mastodon, software created in 2016 by the German developer Eugen Rochko that uses the ActivityPub's affordances to reconstruct several features of Twitter. As such, Mastodon serves as a microblogging

platform where users can post short *toots* (up to 500 characters) and read a feed of *toots* posted by people they are following. Users can interact with one another, liking *toots*, responding to them with their own comments and sharing them on their followers' feeds (*boosts*), increasing the visibility of certain topics. The moderation of these interactions, however, occurs in a decentralised way, on the level of the communities, as explained in the next section.

The rise of Mastodon revisits discussions about openness and decentralisation in the development and governance of digital technologies. These have been core principles of internet ideology, whereby technological decentralisation can lead to political democratisation (Benkler, 2006; Miller & Vaccari, 2020). It is ever clearer, though, that the internet has not evolved towards decentralisation, as corporations and states have leveraged their power and strongly shaped the development of the network. For the FOSS movement, proprietary technologies are to blame, as they allow powerful entities to keep control over digital resources (Couture & Toupin, 2019; Rosnay & Musiani, 2020). Free software, on the other hand, could return control to individuals, social movements and small entrepreneurs (Couture & Toupin, 2019). Even Global South countries have seen in the FOSS movement some potential to counter the power of rich nations over technologies of information and communication (Schoonmaker, 2018; Tomaz, 2025). Although the movement has always tried to distance itself from politicisation, it clearly resembles liberal ideas of power dispersion that have shaped Western democracies, including their media and communication systems (Coleman, 2004).

However, scholars increasingly question the conflation of technological decentralisation with democratisation (Bodó et al., 2021). Indeed, while FOSS and other volunteer-based projects, such as Wikipedia, can offer their participants a level of autonomy greater than the typical experience of the user of mainstream social media platforms, they nevertheless enable concentrations of power and authority (O'Neil, 2009). In addition, the liberal focus on individual freedom often conflicts with non-Western views on democracy, communities and individuals, challenging the universality and neutrality of technological openness (Mansoux & Abbing, 2020, p. 131). Mansoux and Abbing argue that the Fediverse represents a step forward in this sense. As will become clearer in this paper, the ideas of federation and community-oriented governance acknowledge that openness has limits and there is a move towards a more social negotiation of the use of technological resources.

Content moderation on Mastodon

Content moderation, although often unseen, is the very service that social media

platforms offer to their users (Gillespie, 2018). It can be understood as "the process in which platforms shape information exchange and user activity through deciding and filtering what is appropriate according to policies, legal requirements and cultural norms" (Zeng & Kaye, 2022, p. 81). This is as true for decentralised social media as it is for the large commercial platforms. Indeed, volunteer- and peer-based online projects should not be misinterpreted as inherently anarchical, as they often produce governance and policy infrastructures that enable cooperation (Butler et al., 2008). However, in contrast to their commercial counterparts, decentralised social media like Mastodon leverage their technical affordances to offer users a plurality of content moderation, as each instance can adopt its own policies from liberal, light-touch moderation, to stricter forms (Zulli et al., 2020).

On the technical level, Mastodon seeks to facilitate plurality by relying on self-hosting and providing a set of moderation tools to admins and moderators. Mansoux and Abbing (2020) interpret these unique features of Mastodon as socio-technical affordances that move away from previous techno-deterministic attempts of FOSS to embed values in the code itself and acknowledge the agency of humans in interaction with software. Self-hosting means that there is no generic 'Mastodon' a user joins. A new user must always join a so-called *instance*, which is a Mastodon installation provided by independent people or organisations that then connects to the bigger network of existing instances. A user can even create an instance for themselves. This shifts moderation decisions to the level of the instances, which can establish their own rules, such as who is allowed to join their community or not. Some instances are closed groups that only accept new members by invitation. Others accept new members that share similar interests with an approval-only system. A few, usually larger instances approve any new member without conditions. Over time, individual instances also experiment with more than one of these possibilities, often having a loose approach for a while and closing the group when it reaches the limit of moderation or server capacities.

Therefore, in Mastodon, users must abide by rules and policies of their instances as decided by the respective admins of those instances. Federation means, however, that they can still follow people from other instances. If instances find the content permitted on another instance unacceptable, they can block or 'defederate' with one another. The content continues to exist on the more permissive instance, but it can no longer be accessed from the more restrictive one, and its users are protected from the offending content. According to Mansoux and Abbing (2020), this practice resembles a pluralist understanding of democracy, in which the other retains the legitimacy to express their view, but does not necessarily have the right to re-

quire others to hear them. While Mastodon instances are autonomous, it is common for the rules and policies of instances to share similarities, such as prohibiting racism, sexism and inappropriate depictions of children. Thus the federation of Mastodon instances – or the Mastodon-Fediverse – can be described as engaging in "a platform governance model of *covenantal federalism*, where small units consent to band together while abiding by a shared ethical code" (Gehl & Zulli, 2023, p. 3276).

The option to defederate between instances has helped Mastodon manage the most significant challenge to its content moderation so far: the arrival of Gab in 2019, an instance that promoted far-right and white supremacist content. The Mastodon-Fediverse was able to demonstrate effective self-regulation, with many major instances defederating from the interloper, effectively freezing it out of Mastodon's ecosystem (Caelin, 2022; Rozenshtein, 2023). Indeed, Rozenshtein (2023) describes this as a success story of Mastodon's "content-moderation subsidiarity" (p. 228), with the independence of its instances allowing it to maintain equilibrium across the network, without impinging on user choice.

In less extreme cases, admins may silence – instead of block – instances whose content do not follow the community norms, to allow individuals to make their own choice. With this option, individuals can still follow accounts from silenced instances, but will not be able to share that content via boosts with fellow users from their original instance. As an example, Todon.nl, one of the biggest Mastodon instances, explicitly silenced content from the now defunct Switter.at "because a lot of images are still not marked as sensitive" (ghost, 2018). Over time, some users started to curate lists of "bad actors" that can be used by new instances to block or silence (e.g., Fediblock, Garden Fence, The Bad Space, Seirdy's and Oliphant's lists), reducing the burden of moderation.

Instances' admins can also decide who participates in the administration and moderation process and their specific rights, such as accepting new users or deleting content. Liberal, generalist instances can have a small pool of moderators with simple duties, whereas more focused instances usually recruit moderators among active community members who have a good understanding of their rules. For most moderators, this is a volunteer, part-time job. Normally relying on donations, instances are rarely able to remunerate moderators, raising the longstanding issue of free labour in the FOSS movement (Anaobi et al., 2023; Mansoux & Abbing, 2020; Zhang et al., 2024). On the other hand, volunteer moderators have a closer relationship with their communities, giving them a better knowledge of the problems and implications related to their decisions.

Moderators directly intervene on user and content levels. Users can be suspended for a time - with the right to access the account and read their feeds, but no right to toot - or even banned if they do not comply with the community rules. Admins and moderators can also delete posts that infringe their rules. In this sense, they play a similar role to moderation in commercial platforms, but these decisions are always made on the instance level, by their admins and moderators, often in exchange with the respective community of users (Caelin, 2022). Because of the federation system, users unsatisfied with their instance can move to another one, import their list of followed accounts and access their feed in the new home.

In contrast with corporate social media platforms, where users often notice content moderation only when they fall foul of the rules, norms have appeared on Mastodon that require a certain level of active user participation and community engagement that have been born out of the involvement of members of marginalized and often queer communities (see Mansoux & Abbing, 2020). In this sense, content moderation on Mastodon more closely resembles James Grimmelmann's classic description of content moderation as "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" than the centralised, top-down models of corporate social media platforms (Grimmelmann, 2015, p. 42). For example, Mastodon has 'content warnings' (CW), a feature that enables users to hide sensitive content behind a button. Users can write a short title describing their post, such as 'US politics', placing the content behind the CW. Historically, many instances encourage their users to use CW for political content or images with eye contact, which might negatively affect people within the autistic spectrum who may find direct eye contact uncomfortable (see Stewart, n.d.). This can be described as a practice of visibility moderation, whereby not just the availability of content is of importance, but also the context in which it is available (Zeng & Kaye, 2022). Furthermore, while both Mastodon and Twitter/X enable users to provide 'alt' text descriptions of images for the benefit of users with visual impairments, many Mastodon instances and users actively encourage the use of this option. Indeed, on some instances, whenever someone toots an image without alternative description, a bot warns the user that people with visual impairment cannot properly read and suggests that the user re-post the image with alt text.

These are the main characteristics of content moderation on Mastodon, which in many aspects differ from corporate social media platforms. In this article, we group them into five major areas: *user acceptance*, *recruiting moderators*, *federation/defederation*, *direct intervention on user/content level* and *specific norms*. The focus of

the empirical investigation is what happens when these areas face the challenge of sudden growth in user numbers.

Becoming 'popular-by-surprise'

Even as the flagship of the Fediverse, with over 90% of its active users, Mastodon remained a niche phenomenon until 2022, with no more than 300,000 monthly active users, nowhere close to the hundreds of millions to billions of users of commercial social media platforms. However, Mastodon entered the spotlight when Elon Musk bought Twitter and started a series of governance changes blamed for an increase in hate speech and abuse. His controversial management spurred what came to be known as the #TwitterMigration, whereby dissatisfied Twitter users looked for alternative homes for their digital activities. Mastodon became, then, the main destination of "Twitter migrants", gaining more than two million new users in the following two months (Chambers, 2022; Peters, 2022).

As Ermoshina and Musiani (2025) point out, this kind of platform migration is not unprecedented. Mastodon has experienced previous – albeit smaller - spurts of growth in user numbers in the past, particularly in response to unpopular changes to the Twitter/X platform and the Cambridge Analytica data scandal at Facebook (Jeong, 2017; Fung 2018). Indeed, Gab was one of the platforms that benefited from the mass movement of alt-right users who found Twitter/X's content policies to be too strict (Kor-Sins, 2023). Ermoshina and Musiani (2025) identify a wave of "affected activists (from both extremes of the political spectrum), marginalized populations, tech enthusiasts and journalists switching from Twitter/X to decentralized and open source tools that constitute the Fediverse, where Mastodon is an outstanding example" (p. 3). As compared to the effects of deplatforming, in which users are forced to seek alternatives as a result of platforms cracking down on right-wing extremism and conspiracy theories (Rogers, 2020), these users seem to have switched in reaction to an increase in objectionable content on Twitter/X (Cava et al., 2023).

For Mastodon, this posed immediate problems for its servers, many of which were overloaded by new users and became temporarily unavailable (Hoover, 2022). The previous experiences of start-up social media platforms have shown that becoming 'popular-by-surprise' poses particular challenges for content moderation (Gillespie et al., 2020). Companies like Instagram experienced sudden stress early on in their histories, when they were attempting to manage a sudden surge in users with the resources of a small operation (Frier, 2020). In this stage, platforms typically need to increase the number of staff and introduce automation to cope with greater vol-

umes of problematic content, as well as write comprehensive and professionalised community guidelines, where previously rules had been minimal (Caplan, 2018).

At the same time, the sudden growth in Mastodon user numbers was never guaranteed to be sustained. As Cava and colleagues (2023) note, "transitioning to a different social platform entails practical and psychological costs associated with changing habits, as well as a social cost associated with adopting a behavior that deviates from mainstream norms" (p.1). Indeed, while users may have been searching for an alternative to Twitter, they may have experienced a 'culture shock' due to Mastodon's hidden but significant differences. Adapting to Mastodon's pre-existing culture may increase perceived participation costs of "how much time and effort are required to engage with content provided in a community" (Butler et al., 2014, p. 699). Users may also have had to accept the cost of losing their pre-existing networks on Twitter. These switching costs may be behind the slump in the numbers of monthly active users on Mastodon to 1.4 million by the end of January 2023 (Hoover, 2023).

While there is some established understanding of how sudden unanticipated user growth affects content moderation on centralised social media platforms, moderation in Mastodon instances differs in key ways, as explained above. The influx from Twitter can be understood as a stress test for the internet alternatives to large commercial companies. Understanding the challenges that Mastodon instances have faced and how they have met them can help build an understanding of which resources are needed to strengthen the civic internet for a potential future where users turn to in large numbers.

Methodology

To capture continuities and changes in content moderation, we conducted eight semi-structured interviews. Qualitative semi-structured interviews have proven to be an effective method to understand practices, negotiations, key decisions and dilemmas by decision-makers - in this case, administrators and moderators.

Seven of these interviews were conducted with administrators and moderators of Mastodon instances about practices before, during and after the influx from Twitter in late 2022. The chosen instances were Colorid.es, Khiar, Mastodon.uno, Mastodon.uy, Mastodon.world, Openbiblio.social and Ursal.zone. They reflect significant differences in size (large/small user bases), thematic focus (generic/topic-specific) and geography (instances based in the Global North and South, as well as a diversity of languages), and the common thread among them is the sudden

growth in number of registered users after October 2022. All interviews were conducted online with either one admin or moderator of the respective instance, except for the interview with the staff from Mastodon.uy, where the team of three admins and moderator preferred to have the conversation together. Apart from the interview with Mastodon.uno, held in English, all interviews occurred in the main language of the instance (either English, Portuguese, Spanish or German).

At the time of data collection, Mastodon had over 1,100 instances with more than 100 users each. The biggest one was, and still is, the generic Mastodon.social, operated by the software development team, with 1,832,866 registered users in February 2024. Despite its size, it still accounted for only 20.4% of the total number of Mastodon users, keeping the network fairly decentralised. However, there is a Global North-South geographical imbalance, as 71.8% of the users are in English-speaking instances.[1]For this reason, Global South instances remain understudied, hence our interest in a diverse sample. Table 1 presents the instances of our sample as well as their predominant language, user base, a short description and the role of the interviewee.

TABLE 1: Instances in the sample

INSTANCE	LANGUAGE	USERS (2022)	USERS (2023)	DESCRIPTION	ROLE OF INTERVIEWEE
Colorid.es	Portuguese	895	2,230	Oriented to LGBTQIAPN+ Portuguese-speaking people	Mod
Khiar	Persian (Farsi) and English	88	299	Generic instance for Farsi and/or English speakers	Mod
Mastodon.uno	Italian	22,983	69,926	Largest Italian-speaking instance, generic	Admin
Mastodon.uy	Spanish	266	1,355	Generic instance from Uruguay oriented to Spanish-speaking people	Admins and mods (three participants in the same interview)
Mastodon.world	English	121	178,791	10th largest Mastodon instance, generalist, administered from the Netherlands	Admin

1. These figures are based on information provided by https://fediverse.observer/, as of 20 February 2024.

INSTANCE	LANGUAGE	USERS (2022)	USERS (2023)	DESCRIPTION	ROLE OF INTERVIEWEE
Openbiblio.social	German	406	1,300	Oriented to German-speaking library workers, hosted by the Berlin State Library	Admin
Ursal.zone	Portuguese	921	2,796	Brazilian instance focused on anti fascist, left-wing debate	Admin

All figures according to https://fediverse.observer/, comparing Oct 2022 and 2023. Figures refer to the total number of registered users, not "monthly active users", whose records have been inconsistent.

In addition, a further semi-structured interview was conducted with a representative of IFTAS, a nonprofit organisation founded in 2023 to support the volunteer moderator community of the Fediverse, largely as a response to the challenges caused by rising user numbers since October 2022. As IFTAS performed a needs assessment with 134 respondents in August 2023, the interview provided a broader contextual background.

The questions cover the five moderation areas we identify in our theoretical framework and prospects for the future. Our interviews were also informed by an analysis of the "About" pages of each instance, in which we preliminarily looked for information on these topics. In the following, we present our findings.

Findings

New user registrations

For our sample, we selected only instances that experienced significant increases in new user registrations in November 2022. Particularly for generic instances, publicity through media coverage or being listed on Fedifinder while other instances had crashed drove very high rates of new user growth. In the case of Mastodon.uno, Italian media coverage led to 10,000 new users in two days. In our sample, several instances responded to this by restricting registrations, either permanently or temporarily. Mastodon.uy introduced approval-only registrations - where new user applications must be approved by a moderator - to make sure only Spanish-speakers would join the group. Colorid.es put age restrictions and a very specific procedure to avoid disruption: after the 20th request on the same day, admins and moderators of the instance waited 30 minutes before approving the next

user to avoid an excessive number of new users in a short time. Mastodon.world, which has an open registration policy, temporarily closed registrations on multiple occasions while solving problems. For example, it did so while waiting to receive updated Mastodon software that included a reCAPTCHA feature to prevent the registration of spam accounts. These limits to registration have previously been used by many instances in the past (Zulli et al., 2020). However, as our sample shows, several instances maintained more liberal practices until this influx, when they also felt the need to adjust. On top of restrictions, most instances used communication strategies to manage the expectations of the new users. An Ursal admin "published a thread to explain politics, content warnings, federation, alternative image text and so on, but no further changes" (Interview, Ursal.zone). From January 2023 on, the number of active users dramatically decreased. Some instances remained much bigger than before, such as Mastodon.world, Mastodon.uno and Ursal. Others returned to previous levels of user activity.

Content moderators

Despite solely relying on volunteer work, most of the instances were able to expand their moderator teams during the period of user growth. For example, Mastodon.uno increased from five to 10 moderators and Ursal from two to six. Mastodon.uno's interviewee explained that recruiting new moderators was not challenging: "This is the biggest Italian instance. There is prestige involved, so people are happy to moderate it" (Interview, Mastodon.uno). Khiar also had "a prominent user promoted to moderator" without difficulty, evidencing how content moderators of Mastodon instances are closely tied to their user base (Interview, Khiar). New mods were trained informally, mostly because they were already part of their respective communities. Coordination between moderators increased, often with the help of moderators' chats on Telegram or open alternatives, such as Matrix and XMPP. Most instances managed to keep their new moderators, but a few stepped back, especially as the migration wave passed and many new users did not return. In particular, Mastodon.world faced the same level of content moderation workload with fewer moderators than directly after the influx. Furthermore, Mastodon.world noted under-resourcing not just in terms of the number of moderators, but also of the need to find moderators with culturally and geographically diverse backgrounds who are able to understand the context of certain debates and moderate appropriately. Smaller, more niche instances, such as Openbiblio.social and Khiar, feel that their moderator teams are sufficient, because of the limited growth in users and limited changes to content moderation workload.

Defederation practices

In turn, federation and defederation practices do not seem to be impacted by the events of October 2022, and as such, no change was reported. Most instances follow hashtags such as #fediblock or blocklists, but claim to still not adopt mass defederation, rather opting for default federation with all instances in the Fediverse, blocking unwanted instances on a case-by-case basis, in a practice that predates the influx of new users in 2022. In some cases, such as Ursal, admins are also informed by exchange with admins from other instances. For Mastodon.uno, the practices of other Italian instances spark considerations on the matter: "We see what other Italian instances are doing. If they are blocking a bad instance, we start discussing what to do. If we find that they are right, we agree to block. But it's really rare that this happens" (Interview, Mastodon.uno). In general, defederation is not proactive, but reactive: users report problematic content they see on the federated timeline feed and only then do admins and moderators consider defederation or 'silencing' the instance in question. While opting to block undesirable instances on an ad hoc basis, Mastodon.world also worked together with users to identify and block abusive accounts on other instances. In particular, "We've had our largest account on our server, [which is] the account of the Auschwitz Museum, which obviously especially in the beginning drew a lot of people that were like Holocaust deniers and extremists that were harassing that account and then we worked together with that account and they reported them really quickly and we just removed and blocked them really quickly, so that also cleaned up nicely" (Interview, Mastodon.world).

Setting and communicating rules and norms

When it comes to setting and communicating norms of how users should behave, the instances reacted heterogeneously to the influx of new users. While Mastodon.world quickly drew up a set of rules where the only previous rule was 'behave', instances with pre-existing rules did not change them substantially. However, in some cases, there were changes in the enforcement of these rules. In particular, Mastodon.uno became "very flexible about this, because we have a lot of people breaking these rules. We didn't take action, we didn't remove messages for two to three months. It was impossible for us to check every message". Interviewees also reported cultural clashes between instances and new users who were unfamiliar with the culture of Mastodon. Colorid.es has a high adherence to Mastodon-specific norms and complained that new users arrived expecting "a Twitter copy". Ursal also engaged specifically with former Twitter/X users, addressing their alleged tendency to fuel divisive discussions: "There was newbie behaviour

from Twitter users aggressively fighting about everything. We had to explain that you can't make a fight viral [on Mastodon], so the troll is only fed if you keep answering them" (Interview, Ursal.zone). A similar approach can be seen in Mastodon.uy: "[New] users must understand that you have to change your mindset when you use the [FOSS] social networks. Twitter is a space to go and rant about any content. It's different from the Fediverse. [Here] we are concerned about maintaining, guarding the community. With new users [during this influx], it has never been a source of great conflict, but users have to make an effort to adjust to a different form of communication" (Interview, Mastodon.uy). These quotes show that Mastodon norms were often explained in contrast with Twitter/X, suggesting a certain narrative about how the technical affordances of each model favour different kinds of discourses. This influx led, however, to some more specific clashes. IFTAS noted that on some instances, incoming Black users have encountered the pre-existing expectation on Mastodon that all content about politics should be covered with a content warning and thus received a warning for posting about racism and the #BlackLivesMatter movement without one. Moderators from Ursal, Mastodon.uy and Khiar point out that they already had a very liberal approach to content warning for political content, considering this norm rather "depoliticisation" or "Global North hypersensitivity", so they did not try to enforce it at all. A Mastodon.uy admin said: "About content warning, I have the feeling that it is a thing of people from the North [US and Europe]. There is less sensitivity in Latin America. It is only used for very extreme cases" (Interview, Mastodon.uy). In a similar vein, "People in the West and in Europe are a lot more sensitive. They care more about trigger warnings. Iranian people don't care. They can put NSFW for nudity. [Apart from that] they don't really follow these norms [such as CW]" (Interview, Khiar).

Direct interventions

Regarding direct interventions by admins and moderators, such as deletions and account suspensions, most of the instances interviewed experienced a significant increase in reports and therefore a significant increase in interventions after the influx of new users. In exceptional cases, such as Mastodon.uno's, content moderation was temporarily suspended. "At its peak, we had 10 messages published every second from new users, it was impossible to moderate" (Interview, Mastodon.uno). After the number of users started to decrease again, the workload remained high for larger, generic instances as controversial geopolitical events such as the conflict in Gaza increased the potential for disputes and uncivil speech. This presents not just a quantitative, but also a qualitative challenge, as "… obviously a report on a Israel-Gaza-Hamas thing is way more work to read and understand than some-

one just reporting spam or [something] that's really easy, that's click-click and it's done, but this is [something] we need to read, and read the context, and that's a lot more work. So it's not only the number that's increased, but also the work involved that's increasing" (Interview, Mastodon.world). Meanwhile, smaller, niche instances returned to pre-migration levels of content moderation workload. A few instances, however, have never really had an increase in the number of sanctions. This was the case of Openbiblio.social, because it attracts mostly people from a specific community where users are likely to know each other in offline settings, establishing a level of "social control" against bad behaviour. Mastodon.uy also reported no cases of content removal or bans except for one single post with apology to violence.

Prospects for the future

When asked about the future, interviewees felt that, while the influx of 2022 was over, users will continue to turn to Mastodon, albeit at a slower rate. They noted that the number of new user registrations tends to increase in response to negative press coverage and controversial events around Elon Musk. Our interviewees saw growth in numbers of new registrations and user activity levels as a positive development, because it advances the ideal of social media decentralisation and increases diversity in the Fediverse. The interviewee from Mastodon.world expressed a hope that Mastodon would see not just a growth in user numbers, but also a growth in the number of instances. However, IFTAS noted that some members of the Mastodon community would prefer to avoid receiving large numbers of new users to prevent a perceived threat to Mastodon's culture. Mastodon.world also saw the potential for the adoption of the ActivityPub protocol by the Instagram app Threads not just to grow Mastodon's interoperability, but also to benefit Mastodon's profile and visibility. At the time of the interviews, this was being tested and was later implemented, in March 2024.

Discussion

The findings of this study show a notable level of heterogeneity between instances' experiences with content moderation after the influx of new users at the end of 2022. The heterogeneity of experiences can be connected to the heterogeneity of the instances themselves. The representative from IFTAS described three kinds of Mastodon instance: large generic instances without a thematic focus, community-focussed instances with some level of thematic focus around particular shared interests, and specialised managed communities for users with additional safety needs, such as LGBTQIA+ users - although further categories are pos-

sible. Our sample contained instances from all three categories, with large generic instances experiencing the highest level of content moderation stress where workload significantly outstripped resources. Across this diversity of instances, however, three important themes can be highlighted: firstly, the challenge posed by content moderation automation to the ideals of decentralisation; secondly, cultural tensions, particularly in a Global North-South context; and thirdly, the contribution of controversial geopolitical events to content moderation workload.

Firstly, the demand for automated content moderation tools, such as reCAPTCHA and software to block child sexual abuse imagery (CSAM), may pose a challenge to the ideal of decentralisation. As Mastodon instances may not be well-placed to develop or access these resources themselves, they are dependent on external help from developers and organisations within the community, such as the software development team itself and IFTAS (Interview, IFTAS). This can be observed in hosting, such as on Mastohost, and now in content moderation, with the development of tools to combat spam and CSAM. As the representative from IFTAS noted, at the core of this need is the tension between centralisation and decentralisation, as ostensibly sovereign instances still depend on services that are in some way centralised. Mastodon as a network seeks to uphold the values of decentralisation while also refraining from the use of algorithms, for example in recommendation feeds or to target digital advertising. However, the growth challenges of content moderation put pressure on Mastodon instance admins to use tools that are centralised and automated, a trend that appears to run counter to Mastodon's values. In particular, shared content moderation software may pose similar challenges to those posed by the 'content cartels' of large, centralised platforms, where shared automated resources may produce a homogenisation of content moderation across networks (Douek, 2020). At the same time, automation may help to alleviate the workload of under-resourced admins while reducing the exposure of moderators to disturbing CSAM content. This is particularly pertinent in a context where Mastodon instances do not have the financial resources of large social media companies that could fund mental health support to content moderators.

Secondly, the practice of content moderation reveals cultural challenges within the Mastodon-Fediverse that are only made more apparent by the arrival of new users. Instances closer to Global South communities pointed to the fact that Mastodon norms have emerged from Western countries where the majority of Mastodon developers and users are based. These instances have tended to be more flexible regarding these norms and the established Mastodon culture. While they have not shown intense opposition to these norms, their reactions to the sudden influx have

diverged from the established culture. If Mastodon is to grow further, managing this North-South cleavage will require cultural changes and adaptations on both sides. To our understanding, this is already taking place to a certain extent, but there remains a need for the norm around content warnings on political content to be flexibilised, for example to avoid censoring the lived experience of marginalised groups, such as Black users. Here, the tension between centralisation and decentralisation becomes apparent again, as Mastodon technical and cultural development remains concentrated in the Global North and caters to the needs of vulnerable groups as understood within that context, but expansion towards the Global South calls those definitions into question. As such, Global South appropriations of the Fediverse sometimes clash with Western interpretations of democracy, freedom and emancipation, as previously suggested (Mansoux & Abbing, 2020). In other cases, some instances may find benefit in maintaining a certain level of cultural homogeneity. This is underscored by Openbiblio.social that attracts users from a specific professional community and notes that: "Our users obviously read our rules and abide by them. That's good and practical, and perhaps that's really because many of them are colleagues and know each other, and so there's [a] bit of social control offline, off the instance, so it works really well" (Interview, Openbiblio.social). Instances that serve users with higher safety needs may respond by actively taking a more defensive stance to user growth, e.g. by restricting the acceptance of new users. This was the case for Colorid.es, which serves the LGBTQIAPN+ community. The LGBTQIAPN+ community, which has particular safety needs, has played a key role in the early phase of Mastodon and shaped its development in an attempt to provide an alternative to mainstream social media, where they have experienced harassment and hate speech (Valens, 2019). It is notable that Colorid.es, with increased safety needs and identification with Mastodon's established culture, found new user growth more culturally challenging than the other instances.

Thirdly, content moderation workload does not seem to be connected to growth in user numbers alone, but also to the occurrence of controversial geopolitical events, such as armed conflicts and elections. Partly, this could be because of an increase in post volume, but the interviews also pointed to (a) the ethical complexity of content moderation decisions around controversial topics that requires specialist contextual knowledge, and (b) the potential for controversial events to generate online debates that escalate and increase the risk of posts that violate the instance's rules. Even a small, low-workload instance such as Openbiblio.social was not immune to this, while Mastodon.world found this to be a greater challenge than the initial influx itself. If Mastodon instances continue to grow, content moderation of controversial debates will require not just an increase in modera-

tors, but also more diversity in contextual knowledge among moderators that will need to persist even after numbers of monthly active users have abated. These demands are likely to challenge the part-time, volunteer and mostly amateur character of content moderation in federated social media, increasing pressure for professionalisation. While there may be opportunities to learn from other platforms that rely on volunteer content moderators, such as Reddit and Wikipedia, Mastodon differs from these precedents in its level of decentralisation. While both Reddit and Wikipedia have some level of centralisation in their content moderation infrastructure – e.g. Reddit retains the ability to ban subreddits and the Wikimedia Foundation can suspend accounts and pages –, the non-profit that develops the Mastodon protocol does not have the technical ability to suspend instances or ban users. This implies that Mastodon instance admins may experience a higher level of responsibility in exercising content moderation than volunteer moderators on other platforms, which represent opportunities but also challenges, especially when it comes to the complexity of moderating controversial debates.

Mastodon growth has presented several challenges for content moderation. The challenges themselves have been found to be very heterogeneous depending on the type of instance, reflecting the diversity of decentralised social media. Admins and moderators have taken active measures compatible with the kinds of communities they oversee, such as increasing the number of moderators, imposing restrictions to user acceptance or improving the communication of rules and norms. In this sense, federated social media have shown high flexibility and sensitivity to the needs of different communities, which is an advantage over the corporate model. However, tensions remain, such as the need for some centralised resources, the North-South tension and the increasing complexity of the moderation task. Further growth in user numbers is likely to increase these tensions, and it is yet to be seen if the federated model can cope with them to become a real alternative to corporate platforms.

References

Anaobi, I. H., Raman, A., Castro, I., Zia, H. B., Ibosiola, D., & Tyson, G. (2023). Will admins cope? Decentralized moderation in the fediverse. *Proceedings of the ACM Web Conference 2023*, 3109–3120. https://doi.org/10.1145/3543507.3583487

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Bodó, B., Brekke, J. K., & Hoepman, J.-H. (2021). Decentralisation: A multidisciplinary perspective. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1563

Butler, B., Joyce, E., & Pike, J. (2008). Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1101–1110. https://doi.org/10.1145/1357054.1357227

Butler, B. S., Bateman, P. J., Gray, P. H., & Diamant, E. I. (2014). An attraction-selection-attrition theory of online community size and resilience. *MIS Quarterly*, *38*(3), 699–729. https://doi.org/10.25300/MISQ/2014/38.3.04

Butler, B., Sproull, L., Kiesler, S., & Kraut, R. (2007). Community effort in online groups: Who does the work and why? In *Leadership at a Distance* (pp. 187–210). Psychology Press.

Caelin, D. (2022). Decentralized networks vs the trolls. In H. Mahmoudi, M. H. Allen, & K. Seaman (Eds.), *Fundamental challenges to global peace and security: The future of humanity*. Palgrave Macmillan. http://gen.lib.rus.ec/book/index.php?md5=2BE9967A3318C380B758144D48796848

Caplan, R. (2018). *Content or context moderation?* Data & Society. https://datasociety.net/library/content-or-context-moderation/

Cava, L. L., Aiello, L. M., & Tagarelli, A. (2023). Drivers of social influence in the Twitter migration to Mastodon. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-48200-7

Chambers, T. (2022). *A snapshot of the #TwitterMigration*. Dewey Square Group.

Coleman, G. (2004). The political agnosticism of free and open source software and the inadvertent politics of contrast. *Anthropological Quarterly*, *77*(3), 507–519. https://doi.org/10.1353/anq.2004.0035

Couture, S., & Toupin, S. (2019). What does the notion of "sovereignty" mean when referring to the digital? *New Media & Society*, *21*(10), 2305–2322. https://doi.org/10.1177/1461444819865984

Douek, E. (2020). *The rise of content cartels*. Knight First Amendment Institute at Columbia University. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3572309

Dulong De Rosnay, M., & Musiani, F. (2020). Alternatives for the internet: A journey into decentralised network architectures and information commons. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 622–629. https://doi.org/10.31269/triplec.v18i2.1201

Ermoshina, K., & Musiani, F. (2025). Safer spaces by design? Federated socio-technical architectures in content moderation. Internet Policy Review, 14(1). https://policyreview.info/articles/analysis/safer-spaces-design

Farrell, H. (2022, November 12). Musk is wrecking speech moderation on Twitter. There's an alternative. *The Washington Post*. https://www.washingtonpost.com/politics/2022/11/12/musk-is-wrecking-speech-moderation-twitter-theres-an-alternative/

Flensburg, S., & Lai, S. S. (2020). Comparing digital communication systems: An empirical framework for analysing the political economy of digital infrastructures. *Nordicom Review*, *41*(2), 127–145. https://doi.org/10.2478/nor-2020-0019

Frier, S. (2020). *No filter: The inside story of Instagram*. Simon & Schuster.

Fung, B. (2018, March 23). The new technology that aspires to #DeleteFacebook for good. *The Washington Post*. https://www.washingtonpost.com/news/the-switch/wp/2018/03/23/the-new-technology-that-aspires-to-deletefacebook-for-good/

Gehl, R. (2015). FCJ-190 Building a better Twitter: A study of the Twitter alternatives GNU social, Quitter, rstat.us, and Twister. *The Fibreculture Journal*, *26*, 60–86. https://doi.org/10.15307/fcj.26.190.2015

Gehl, R. W., & Zulli, D. (2023). The digital covenant: Non-centralized platform governance on the Mastodon social network. *Information, Communication & Society*, *26*(16), 3275–3291. https://doi.org/10.1080/1369118X.2022.2147400

ghost. (2018, April 16). *Only silence instance on federated timeline (default or option) #7153*. GitHub. https://github.com/mastodon/mastodon/issues/7153

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A., & Myers West, S. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1512

Griffin, R. (2023). Public and private power in social media governance: Multistakeholderism, the rule of law and democratic accountability. *Transnational Legal Theory*, *14*(1), 46–89. https://doi.org/10.1080/20414005.2023.2203538

Grimmelmann, J. (2017). *The virtues of moderation*. LawArXiv. https://doi.org/10.31228/osf.io/qwxf5

Hoover, A. (2022, November 9). *Twitter users have caused a Mastodon meltdown*. Wired. https://www.wired.co.uk/article/twitter-users-mastodon-meltdown

Hoover, A. (2023, February 7). *The Mastodon bump is now a slump*. Wired. https://www.wired.com/story/the-mastodon-bump-is-now-a-slump/

Jeong, S. (2017, April 4). *Mastodon is like Twitter without Nazis, so why are we not using it?* Vice. https://www.vice.com/en/article/mastodon-is-like-twitter-without-nazis-so-why-are-we-not-using-it/

Klemens, B. (2023, January 2). *Mastodon—And the pros and cons of moving beyond big tech gatekeepers*. Ars Technica. https://arstechnica.com/gadgets/2023/01/mastodon-highlights-pros-and-cons-of-moving-beyond-big-tech-gatekeepers/

Kor-Sins, R. (2023). The alt-right digital migration: A heterogeneous engineering approach to social media platform branding. *New Media & Society*, *25*(9), 2321–2338. https://doi.org/10.1177/14614448211038810

Mansoux, A., & Abbing, R. R. (2020). Seven theses on fediverse and the becoming of FLOSS. In K. Gansing & I. Luchs (Eds.), *The eternal network: The ends and becomings of network culture* (pp. 124–140). https://networkcultures.org/blog/publication/the-eternal-network/

Miller, M. L., & Vaccari, C. (2020). Digital threats to democracy: Comparative lessons and possible remedies. *The International Journal of Press/Politics*, *25*(3), 333–356. https://doi.org/10.1177/1940161220922323

O'Neil, M. (2009). *Cyberchiefs: Autonomy and authority in online tribes*. Pluto Press.

Peters, J. (2022, December 20). *More than two million users have flocked to Mastodon since Elon Musk took over Twitter*. The Verge. https://www.theverge.com/2022/12/20/23518325/mastodon-monthly-active-users-twitter-elon-musk

Pierce, D. (2024, February 7). *The fediverse, explained*. The Verge. https://www.theverge.com/24063290/fediverse-explained-activitypub-social-media-open-protocol

Rahman, K. S., & Teachout, Z. (2020). *From private bads to public goods: Adapting public utility regulation for informational infrastructure*. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/from-private-bads-to-public-goods-adapting-public-utility-regulation-for-informational-infrastructure

Rogers, R. (2020). Deplatforming: Following extreme internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. https://doi.org/10.1177/0267323120922066

Rozenshtein, A. Z. (2023). Moderating the fediverse: Content moderation on distributed social media symposium: Media and society after technological disruption: Panel on platform governance. *Journal of Free Speech Law*, *3*(1), 217–236.

Schoonmaker, S. (2018). *Free software, the internet, and global communities of resistance* (1st ed.). Routledge. https://doi.org/10.4324/9781315672786

Stewart, R. (n.d.). *Should we insist on eye contact with people who have autism spectrum disorders*. Indiana Institute on Disability and Community. https://www.iidc.indiana.edu/irca/articles/should-we-insist-on-eye-contact-with-people-who-have-autism-spectrum-disorders.html

Terry, M., Kay, M., & Lafreniere, B. (2010). Perceptions and practices of usability in the free/open source software (FoSS) community. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 999–1008. https://dl.acm.org/doi/10.1145/1753326.1753476

Tomaz, T. (2025). Brazilian activism in Mastodon: Sovereignty discourses between cyberlibertarianism and state-centrism. In M. Jiang & L. Belli (Eds.), *Digital sovereignty in the BRICS countries* (pp. 190–213). Cambridge University Press.

Valens, A. (2019, January 18). *Mastodon is crumbling—And many blame its creator*. The Daily Dot. https://www.dailydot.com/debug/mastodon-fediverse-eugen-rochko/

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to *visibility moderation*: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287

Zhang, Z., Zhao, J., Wang, G., Johnston, S. K., Chalhoub, G., Ross, T., Liu, D., Tinsman, C., Zhao, R., Kleek, M., & Shadbolt, N. (2024). Trouble in paradise? Understanding Mastodon admin's motivations, experiences, and challenges running decentralised social media. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1–24.

Zulli, D., Liu, M., & Gehl, R. (2020). Rethinking the "social" in "social media": Insights into topology, abstraction, and scale on the Mastodon social network. *New Media & Society*, *22*(7), 1188–1205. https://doi.org/10.1177/1461444820912533

Published by

Alexander von Humboldt Institute for Internet and Society

CAIS Research for the Digital Age

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies