Safer spaces by design? Federated socio-technical architectures in content moderation

Ksenia Ermoshina, Francesca Musiani

INTERNET POLICY REVIEW

Volume 14 **Issue** 1

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Safer spaces by design? Federated socio-technical architectures in content moderation

Ksenia Ermoshina *National Centre for Scientific Research (CNRS)*
Francesca Musiani *National Centre for Scientific Research (CNRS)*
francesca.musiani@cnrs.fr

DOI: https://doi.org/10.14763/2025.1.1827

Published: 31 March 2025
Received: 18 March 2024 **Accepted:** 21 October 2024

Funding: The authors gratefully acknowledge the past support of the European project NEXTLEAP (nextleap.eu, 2016-2018) first, then of the ResisTIC project (resistic.fr, 2018-2022), funded by the French National Agency for Research (ANR), and the current support of the ANR project DIGISOV (https://digisov.org/, 2024-2027).
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Ermoshina, K., & Musiani, F. (2025). Safer spaces by design? Federated socio-technical architectures in content moderation. *Internet Policy Review*, *14*(1). https://doi.org/10.14763/2025.1.1827

Keywords: Federation, Content moderation, Censorship, Governance, secure messaging

Abstract: Users of secure messaging tools, especially in communities attuned to the risks of state-based and other forms of censorship, increasingly hesitate to delegate their data to centralised platforms, endowed with substantial power to filter content and block user profiles. This article analyses the role that informational architectures and infrastructures in federated social media platforms play in content moderation processes. Alongside privacy by design, the article asks, is it possible to speak of online "safe(r) spaces by design"? And what is the specific role that human moderators play in federated environments? The article argues that federation can pave the way for novel practices in content moderation governance, merging community organising, information distribution and alternative techno-social instruments to deal with online harassment, hate speech or disinformation; however, this alternative also presents a number of pitfalls and potential difficulties that we examine to provide a complete picture of the potential of federated models.

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

Edward Snowden's 2013 revelations (see Snowden, 2019) have been a landmark event in the development of the field of secure communications. Encryption of communications at a large scale and in a usable manner has become a matter of public concern, with a new cryptographic imaginary taking hold, one which sees encryption as a necessary precondition for the formation of networked publics (Myers West, 2018). Alongside the turning of encryption into a fully-fledged political issue, the Snowden revelations have catalysed long-standing debates within the field of secure messaging protocols. Communities of cryptography developers (in particular, academic and free software collectives) have renewed their efforts to create next-generation secure messaging protocols in order to overcome the limits of existing protocols. Developers and technologists worldwide have a core common objective of creating tools that "conceal for freedom" while differing in their targeted user publics, the underlying values and business models, and, last but not least, their technical architectures (Ermoshina & Musiani, 2022).

This experimentation with different technical architectures has a counterpart in the growing mistrust expressed by users of secure messaging tools towards centralised and proprietary messengers and social media platforms (Ermoshina & Musiani, 2022), and the need to look for alternatives, both socio-technical and political. This adds to the well-documented mistrust towards representative democracies and critique of traditional forms of political participation (Rosanvallon & Goldhammer, 2008; Bennett et al., 2013; Blondiaux, 2017). Furthermore, these two levels of distrust, while they are grounded in mostly separate sets of phenomena, are perceived in some countries as related, due to the "dangerous liaisons" that have been documented in recent history between governments and companies (Musiani, 2013), especially when it comes to online surveillance and privacy (Snowden, 2019). Indeed, users become more skeptical about delegating their data to centralized platforms, endowed, "by design" and "by business model", with substantial power to filter content and block user profiles. In addition to government-imposed internet censorship, platform-based and intermediary-based censorship (Zuckerman, 2010) may affect a variety of user groups, from those who could be classified as far-right to human rights defenders, LGBTQI+ activists or even journalists touching upon controversial topics (see DeNardis & Hackl, 2015).

In this search for alternatives, so-called "federated" architectures as the basis of secure messaging and networking are currently experiencing a phase of increased development and use. They are presented as alternatives, on the one hand, to centralized applications that introduce a 'single point of failure' in the network and lack interoperability, and on the other hand, to the peer-to-peer applications that necessitate higher levels of engagement, expertise and responsibility from the user (and her device). Federation is sometimes described as an ambitious techno-political project; federated architectures open up the 'core-set' of protocol designers and involve a new kind of actor, the system administrator, responsible for maintaining the cluster of servers that are necessary for federated networks. Federation is believed to help alleviating the very high degree of personal responsibility held by a centralised service provider, while at the same time distributing this responsibility and the "means of computing"[1]-- the material and logistical resources needed by the system -- with different possible degrees of engagement, favouring the freedom of users to choose between different solutions and servers according to their particular needs and sets of values.

Rather than focusing on the more "traditional" online content governance question of whether censoring some of these users is legitimate or not, our paper focuses on how specific choices in the creation and deployment of informational architectures and infrastructures of federated social media platforms co-shape technical affordances that inform content moderation processes. Alongside privacy by design (see Cavoukian, 2012), can we speak of online "safe spaces by design" to describe socio-technical arrangements enabling a safer conversation online? And what is the specific role that human moderators play in federated environments?

The article first outlines the theoretical foundations subtending the article as well as the empirical material that is at its core, and the methodology used to analyse it. The article analyses the Fediverse as an alternative model for content distribution and moderation, describing briefly its founding principles and key projects. It then moves to providing an analytical portrait of three case studies that are emblematic of federated architecture-based practices in content moderation governance; these case studies introduce elements of community organising, information distribution and alternative techno-social instruments to deal with online harassment, hate speech or disinformation. In addition to the opportunities that these instruments provide to both instance administrators and "end users," the case studies also reveal the potential difficulties entailed by the deployment of federated models.

1. https://www.chapsterhood.com/2019/03/09/decentralize-or-perish/

Context, theoretical foundations and methodology

In our previous research focused on post-soviet activist and journalist communities and their usage of social media, we have examined an interesting pattern which we have called "digital migration", and that can be likened to "platform switching" as described in management literature (see e.g. Tucker, 2019). At least two important waves of migration were identified: Vk.com to Facebook (2011-2012) and Facebook to Telegram (2016-2018). Nowadays, due to recent controversies around Telegram's potential collaboration with the Russian government (Ermoshina & Musiani, 2021) a third wave of migration has been initiated, which involves adoption of decentralized alternatives (Matrix/Element, Mastodon, Pleroma, Delta.Chat, etc.). The context of war in Ukraine and subsequent information control practices have provided further opportunities for federated open source platforms to appear as a possible alternative, offering reliability and resistance to censorship.

In the so-called "Global North", a similar migration wave affected activists (from both extremes of the political spectrum), marginalized populations, tech enthusiasts and journalists switching from X/Twitter to decentralized and open source tools that constitute the Fediverse, where Mastodon is an outstanding example[2]. Now counting several million active users, this platform proposes a federated infrastructure for microblogging and has been hailed as an example of "democratic digital commons" (Kwet, 2020).

Adding to the nascent scholarship that investigates the interplay of federated architectures and content moderation governance (Hassan, 2021; Gehl & Zulli, 2023; Rozenshtein, 2023), this article argues that federation can pave the way for novel practices in content moderation governance. By merging community organising, information distribution and alternative techno-social instruments to deal with online harassment, hate speech or disinformation, federation offers a model that relies on a multitude of "safer spaces". However, this alternative also presents a number of pitfalls and potential difficulties that need to be examined to provide a complete picture of the potential of federated models.

The term "safe**r** space" as opposed to "safe space" is borrowed from an interview we conducted on 19 April 2017, during our fieldwork with a Russian feminist activist, L. She critically assessed the techno-optimist promise of absolute safety and

2. After the election of Donald Trump and his inauguration in January 2025, an important number of users belonging to such groups, as well as research and academia, have also migrated to Bluesky. See Mallapaty, S. (2024). 'A place of joy': why scientists are joining the rush to Bluesky, Nature, 21 November, https://www.nature.com/articles/d41586-024-03784-6

privacy online, arguing that any online platform, even the most private, is potentially vulnerable to hate speech, and that decentralization offers only partial protection against it.

The notion of "safer spaces" has been recently conceptualised in relation to both offline and online contexts that are "framed by a series of boundaries, principles and practices"; the feeling of safety is achieved precisely by the act of voluntary segregation that is sometimes even qualified as "separatism" (Deller, 2019). We argue that federated platforms provide affordances for self-defined communities to set up and maintain their own safer spaces that are united by specific sets of values, topics of interest or cultures.

Taking federated architectures as our core analytical object, we pay particular attention to their interfaces and the underlying protocols of these tools (for example, the core role played by the ActivityPub protocol and the interoperability it offers). Understanding information architectures from a perspective informed by science and technology studies (STS), from foundational contributions to infrastructure studies (Star, 1999) to more recent contributions in software studies (Fuller, 2008), we analyse software as co-producing particular forms of participation. We argue that protocol and interface properties of these federated platforms can diminish possibilities for disinformation, surveillance and online harassment, compared to centralised platforms such as X and Facebook. We will focus on content moderation practices embedded in the architecture of federated tools, but also show the limits of the "safer space by design" approach and the decisive role of community, politics and human agency. The empirical part of the article is organized around three case studies: federated microblogging/social networking service Mastodon, federated real-time communications protocol Matrix.org, and the controversy surrounding Meta's launch of its microblogging platform Threads in May 2023.

Mastodon and Matrix are the two most widely-used cases of federated applications that share several features, such as their decentralised architecture, the possibility of self-hosting, open source code and interoperability. Mastodon is a social network while Matrix is an end-to-end encrypted messaging service. Comparing these tools helps us to explore the effects of federation, as a type of architecture, on content moderation practices.

Within this approach, grounded in STS and more specifically in infrastructure studies, which allows us to pay particular attention to the architectural and infrastructural aspects of federated platforms, this research situates itself in a dialogue with the very recent and burgeoning literature that seeks to examine the questions of

content moderation and governance of the Fediverse (Bono et al., 2024); in addition to the previously mentioned works, already published in peer-reviewed journals, we wish to mention the interdisciplinary works-in-progress presented as the June 2023 "Mastodon Research Symposium" at the University of Warwick[3]. These recent efforts to theorise the Fediverse "from the inside", by practitioners, instance administrators or active users, show an ecosystem in-the-making, and an important degree of reflexivity of the Fediverse communities upon themselves, their tools and their practices.

This article relies on an online-ethnographic study of the federated secure messaging and microblogging platforms Mastodon and Matrix, and the related user communities. The study has included interviews with users (20) and developers (5) of federated messaging applications and a follow-up series of interviews with the Fediverse server or instance administrators (seven interviews), and periods of online ethnography of discussion fora for developers and moderators of federated tools (e.g. the Social Web Incubator Community Group of the W3C[4]). This research was initially conducted in the frame of an H2020 interdisciplinary project on decentralised encrypted messaging, NEXTLEAP ("NEXT generation techno-social, Legal Encryption, Access and Privacy", nextleap.eu, 2016-2018) and has been continued independently by the authors since the official end of the project (see e.g. Ermoshina & Musiani, 2021 and 2022).

The rise of the Fediverse within broader "digital migrations"

During the above-mentioned NEXTLEAP project, we conducted a previous study (2016-2018) with 90+ users of end-to-end encrypted messaging applications. In this study (Ermoshina & Musiani, 2022), we explored (besides other research questions) the motivations behind user preferences for a particular secure messaging application. In the context of a vibrating market of "privacy by design" apps, why do users trust one tool more than the other? For the majority of our user interviewees, the choice was not based on the cryptographic properties of a messenger; on the contrary, even the so-called tech savvy users (developers, cryptographers, digital security trainers) often opted for a less secure tool even though they knew it had security flaws (Ermoshina & Musiani, 2022, pp. 66-88). For instance, the suc-

3. Mastodon Research Symposium, June 22, 2023, University of Warwick, UK and online https://warwick.ac.uk/fac/cross_fac/cdi/news-events?newsItem=8a1785d787044e9c0187045657fe000a

4. Website of the Social Web Incubator Community Group of the World Wide Web Consortium (W3C), https://www.w3.org/community/socialcg/

cess of Telegram in Russia, that we thoroughly analysed in a dedicated paper (Ermoshina & Musiani, 2021), had very little to do with the quality of the actual cryptographic protocols used by Telegram, which are largely criticised by the security community (e.g. Albrecht et al., 2022). Instead, the choice of Telegram was for many users based on the apps' branding, its charismatic leader and the relative openness of its API. This made Telegram attractive for the community contributors to build bots, create stickers or develop independent forks of the app.

However, our analysis also showed that platforms and tools have popularity trajectories: they experience heydays and declines, and user trust should not be taken for granted. Several waves of "digital migrations", as described above – transitions of users from one platform to another in reaction to a specific event, often technical or political – have taken place since the early 2010s. Thus, Snowden's revelations played a crucial role in users' migration from the unencrypted Facebook Messenger to end-to-end encrypted tools such as Signal. Conversely, the unban of the (end-to-end encrypted, but heavily criticised from a technical standpoint) Telegram in Russia in June 2020, and the recent decision by Pavel Durov, its creator, to collaborate with several governments for lawful interception (Germany, for instance) led to waves of migration of users from Telegram to Matrix, Delta Chat or Jabber[5]. Other reasons for waves of digital migration can be connected to changes in the legislation of a country or even shifts in a tool's business model and leadership. For example, when Pavel Durov, after a considerable amount of pressure from the authorities, sold the 'made in Russia' social network Vkontakte or Vk.com to the Russian oligarch Alisher Usmanov, the platform became not only much more commercial, but also open to direct collaboration with the police, which led to a mass migration from Vk to Facebook (see e.g. Butcher, 2014).

It is noteworthy that digital migration is not a linear process; it is not always unilateral and not always exclusive. A user can be co-present in multiple online worlds, and navigate in a "multi-tool setting" as their online personas and threat models are intrinsically multiple (Casilli, 2015; Ermoshina & Musiani, 2018). Users may be present on both Telegram and Signal, or on Twitter and Mastodon, and often cross-post on several platforms manually or using automated solutions (bots or bridges), in order to negotiate parts of their online identity as well as multiply their online presence, and address different target groups associated with those platforms, contributing to several distinct technocultures.

5. See e.g. this discussion on the forum of OpenStreetMap France: https://forum.openstreetmap.fr/t/passage-de-groupes-telegram-vers-signal/25349/5

One of the most striking examples of this migration process is linked to the rise of the Fediverse, an umbrella concept that "refers collectively to the protocols, servers, and applications" (Rozenshtein, 2023) that enable federated social media. The backbone of Fediverse is ActivityPub, a protocol that can be used for sharing different kinds of social media content, from text to photo and video, which makes various services within Fediverse interoperable. Fediverse offers alternatives to the most popular social platforms: Facebook, Twitter, Instagram, Youtube, suggesting open source and federated equivalents (e.g. Frendica, Pleroma and Mastodon for social networking and microblogging, Pixelfed for image sharing, Peertube for video streaming). All of these services can "talk to one another", and potentially respond to users' needs for plurality of tools and content forms.

In our previous work (Ermoshina & Musiani, 2022, pp. 66-88) we have analysed federation as both an infrastructural and a social experiment where developers seek to achieve a compromise between high levels of security and better usability, tackling preoccupations that are 'ideological' and pragmatic at once, e.g. distributing responsibilities among stakeholders, offering particular versions of online freedom, or giving users the choice of their level of autonomy in the system. We have proposed a framework called the "four Cs of federation" for systematising and conceptualising federation. The "four Cs" include:

1. community – (self)-governance and advancement of federated projects depends on engaging a variety of service providers and clients into accepting new open protocols or new libraries; communication and consensus among various projects are needed;
2. compatibility – need to enrol an important number of developers in order to implement and spread particular solutions and being able to secure users; this includes the 'backwards compatibility' needed to enable a harmonious transition from older to more recent protocols without blocking or boycotting 'by design' some of the clients;
3. customisation – possibility to manage smaller user groups and localised implementations, adapting them to the needs of specific user communities without losing the ability to interact with broader networks, which however becomes challenging if there are failures to document systematically all the various implementations of a given protocol;
4. care – the stability of federated ecosystems depends on the successful enrolment of maintainers, which requires the development of good documentation and guides with 'best practices', and the dissemination of technical expertise for future sysadmins.

These "four Cs" will be re-examined in the conclusions of the article in light of its more specific focus on content moderation.

The article will now turn to the empirical core of this study, discussing three case studies: federated moderation governance on Mastodon; the "protocol neutrality" implemented by Matrix.org; the controversy surrounding Meta's microblogging platform Threads. The common thread running through these three case studies is the following research question: by implementing and maintaining particular content moderation arrangements, to what extent are the creators and administrators of federated platforms contributing to build "safer spaces by design"?

Case study 1: Mastodon, or the challenges of federated moderation

The federated microblogging platform Mastodon was launched in October 2016 by Eugen Rochko, a then-24-year-old German developer. However, the tool was relatively unpopular for the first 6 months of its existence, with only around 20,000 users. The first massive migration happened all of a sudden in April 2017, when in two weeks, the number grew up to 365,000 users. One of the reasons for this migration was the controversial US legal bill SESTA (Stop Enabling Sex Traffickers Act) which enabled suspension of sex workers' Twitter accounts (Davisson & Alati, 2024). Another reason expressed by one of our interviewees was *"the rise of hate speech in Twitter from the Trump supporters and all of the hype around fake news, when no one could trust no one anymore"* (interview, Austrian Mastodon instance[6]administrator). At that time, Mastodon enjoyed a lot of media attention, and in a few weeks the first Mastodon instance created by Rochko (Mastodon.social) was full and closed for new users. New instances started to grow fast, which led to some governance-related issues that are not specific to Mastodon *per se*, but are frequent in federated communication services: namely, the question of attributing and enforcing responsibility for user content, and exercising control of the multiple forks and implementations.

The Mastodon ecosystem's core governance tenet is that instances are run by individuals or associations and users are connected to the instance administrators in much more direct and personal ways than it is on Twitter.

> *"Users can 'vote with their feet' by leaving one instance and joining the other, if they are unsatisfied by the way it is run. Or they can take part in the life of the instance, suggest improvements, even ask for changes of some technical parameters, like the*

6. For more details on what a Mastodon instance is and how it works: https://medium.com/@jim-pjorps/a-non-computer-persons-guide-to-how-mastodon-instances-work-da6ceac1994a

> *number of characters that are allowed in a post"* (interview, Russian Mastodon instance administrator)

The functioning of Mastodon instances relies on several layers, from the ActivityPub protocol, the server infrastructure and the software code, on to the Code of Conduct which regulates the behaviour of the users of a particular instance, its values, fields of interest, acceptable and unacceptable content.

Hailed by some as the "Nazi-free Twitter,"[7] Mastodon was promising "safer spaces" to its users via manually regulated, and sometimes almost semi-private, instances. This offered a relatively transparent governance model, with moderators being accessible and responsive to users. However, this changed in 2019, after GoDaddy, Apple and Google banned the right-wing microblogging platform Gab. Gab abandoned its own code and opted for usage of the Mastodon source code, which led to one of the first political statements[8] from the Mastodon core team condemning the usage of their source code by right-wing individuals and collectives as a way to circumvent bans from tech giants. Ultimately, Rochko accepted[9] that he did not have any control over the situation because of the federated nature of Mastodon. The Mastodon community, however, found a way to react to this misuse of their platform, embedded in the very architecture of Mastodon: the right-wing instances were blocked by DNS by many Mastodon instances, therefore isolated or "unfederated". A special account "Isolate Gab" was created and hashtags circulated to demand as many admins as possible to join the "isolation" flashmob. Nonetheless, this response was not propagated to the core of Mastodon's code and rather implemented ad hoc by individual instance administrators[10]. It is noteworthy that, by consequence, Gab's developers have modified Mastodon source code[11] to remove ActivityPub compatibility and possibilities to federate with Mastodon. This was described as a "victory" by the administrators and users involved in the #isolategab

7. O'Neil, L. (2018, August 22). *Tired of Nazis in Your Twitter Mentions? Try Mastodon*. Esquire. https://www.esquire.com/lifestyle/a22777589/what-is-mastodon-twitter-platform
8. *Statement on Gab's fork of Mastodon*. (n.d.). Official Mastodon Blog. https://blog.joinmastodon.org/2019/07/statement-on-gabs-fork-of-mastodon/
9. Robertson, A. (2019, July 12). How the biggest decentralized social network is dealing with its Nazi problem. The Verge. https://www.theverge.com/2019/7/12/20691957/mastodon-decentralized-social-network-gab-migration-fediverse-app-blocking
10. See this discussion of a decision whether or not to modify core in response to Gab switching to Mastodon: https://github.com/mastodon/mastodon/issues/11129
11. *Afaik, Gab doesn't use Mastodon anymore. They did start as a Mastodon instance, ... | Hacker News*. (2021). Ycombinator.com. https://news.ycombinator.com/item?id=25714010

movement[12].

Mastodon's federated architecture actually offers users a different experience as compared to X (formerly Twitter). The user has many options (for instance, to create specific filters for the content that they do not want to see in their feed). The feeds are multilayered, since they can feature not only the "toots" (the Mastodon equivalent of "tweets") published by users of their local instance, but also other instances that their instance is "federating". "Unfederation" is comparable to "unfollowing" on X, but on the level of a server and is usually a decision taken by an instance administrator together with its user community.

Federated social networks introduce novel forms of content moderation, reputation, infrastructure maintenance and community involvement. While in Facebook, the moderator to user ratio was estimated to be 7,500 moderators for 2 billion users, in Mastodon it could be 1 to 500 on some instances, but 1 to 5,000 on others (see Lawson, 2018). And while in the first case manual moderation and user-generated reports of undesirable content can be enough, in the second case it requires optimization. The moderation problem is therefore related to the unexpected fast growth of particular instances, leading to social centralization and lack of capacity of the few (or sometimes the only) moderators:

> "*As a moderator, I might get an email notifying me of a new report while I'm on vacation, on my phone, using a 3G connection somewhere in the countryside, and I might try to resolve the report using a tiny screen with my fumbly human fingers. Or I might get the report when I'm asleep, so I can't even resolve it for another 8 hours*" (Nolan Lawson, Mastodon instance administrator)

12. *IsolateGab :mastodon: (@isolategab@todon.nl)*. (2025). Todon.nl. https://todon.nl/@isolategab/105362599835139257

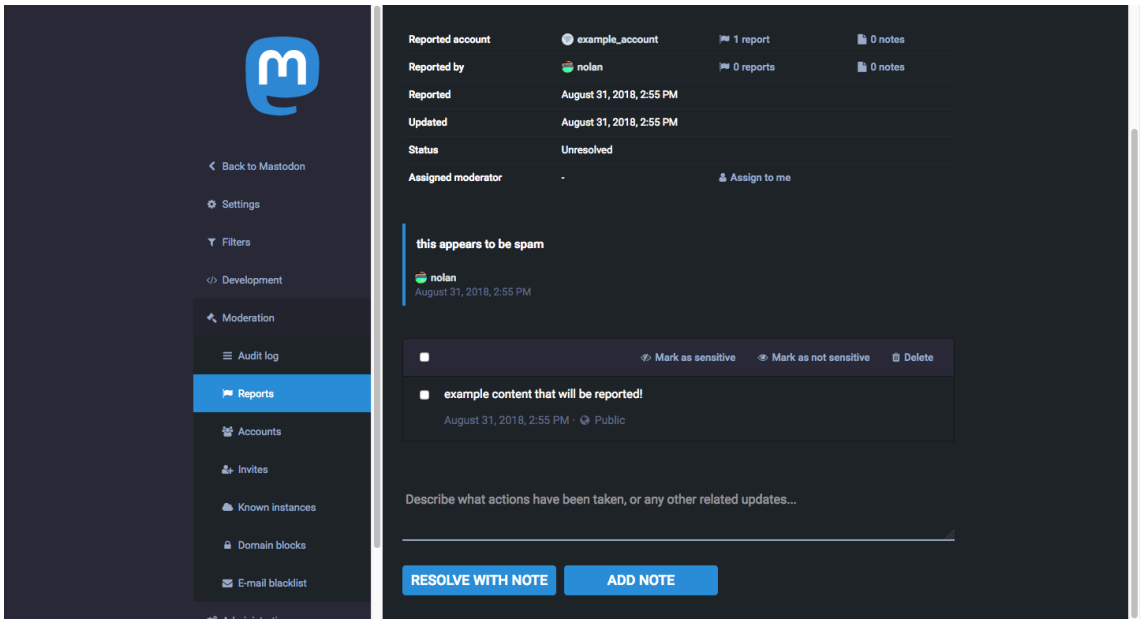

FIGURE 1: The moderator interface for handling reports in a Mastodon instance.

One of the attempts to address this challenge is to automate moderation through the development of bots. Another moderation strategy consists in building relative reputation systems and decentralised identity verification. Relative reputation systems are those that "differ based on the user's position in the network"; such systems allow anyone in the network to "produce subjective scores on network entities or content, published as a reputation feed. Users can combine these feeds in any way to produce their own reputation scoring system" (Graber, 2021, n.p.). This presumes that, unlike in Twitter or Facebook, Mastodon does not push for ID check or any kind of personal data verification; phone numbers or real names are not required.

Finally, one of the most recent suggestions for Mastodon moderation is machine learning. Mastodon's founder has called for ideas about machine-learning based solutions to content moderation challenges; however, the Fediverse community has expressed their skepticism regarding all kinds of automated moderation tools. If ever there are any, they should be instance-specific, and not cross-instance, otherwise it would re-create centralisation; but the implications of this vis-à-vis machine learning is that the learning datasets risk to be rather small, thus reducing the value and validity of these approaches. Moreover, the existing Mastodon moderation documentation clearly stipulates that "moderation in Mastodon is always applied locally, i.e. as seen from the particular server. An admin or moderator on one server cannot affect a user on another server, they can only affect the local copy on their own server"[13].

Therefore, community-driven ad-hoc moderation still seems to be preferred to any "by design" moderation features: with such an approach, a user is asked a standard question about his or her motivations when wanting to join an instance. Some administrators among our interviewees still think this approach is ultimately the best tool to moderate an instance; to us, it seems both ironic and very interesting that the administrators of a platform that is born with a strong connotation of providing a moderation solution embedded in a particular model of network architecture ultimately resort to a very "qualitative" and "human" procedure.

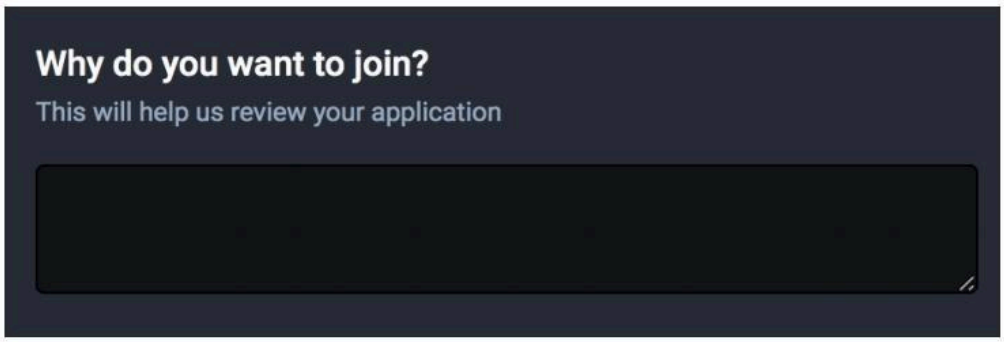

FIGURE 2: Standard question asked to a user wishing to join a Mastodon instance.

Our analysis shows that, while technical decentralization surely enables certain automated practices of content moderation, both the instance administrators and the active user base are deeply involved in decision-making about the Fediverse governance. This includes developing and maintaining codes of conduct for every instance, making key decisions about bridging or not with other instances/servers. Furthermore, the decisions are often based on values subtending those instances.

Moderation concerns are often discussed at dedicated online conferences where instance administrators can take important decisions about the future of Fediverse, such as, for example, an online forum on Mastodon governance and moderation that took place after the "affluence" of far-right users into Mastodon following Trump's expulsion from Twitter[14]. Our interviews with instance moderators and active users, as well as desk research mapping debates on content moderation in Mastodon have enabled us to analyse actual content moderation practices and the

13. *Moderation actions - Mastodon documentation*. (n.d.). Docs.joinmastodon.org. https://docs.joinmastodon.org/admin/moderation/

14. See for example: https://socialhub.activitypub.rocks/t/2021-01-23-socialcg-meeting-new-fediverse-users/1305

role of technology on one hand, and community on the other, in keeping Mastodon's reputation of an online "safer space".

Indeed, large-scale harassment attack is possible in a lot of contexts beyond the Fediverse; however, it "is arguably easier (there) than in a centralized system like Twitter or Facebook, where automated tools can help moderators to catch dogpiling as it happens", as Nolan Lawson, a notorious Mastodon instance administrator stated in his blogpost in 2018, opening a discussion about paradoxes of moderation in Mastodon[15].

On the one hand, this federated microblogging platform suffered from social centralization depending on a small group of admins and moderators, an aspect which was highlighted in our interviews with instance owners as well; for instance, an admin of a Russian instance specifically complained that he could not keep on maintaining it because he was alone. The instance is now discontinued. A solution proposed on the online forum of the Social Web Incubator W3[16]was to limit the size of the instances on the level of all Fediverse, thus reducing the admin to user ratio and supposedly helping moderators to lower the load. However, this kind of centralized (Fediverse-level) decisions are actively criticized in our interviews as "affecting Fediverse freedom".

On the other hand, the report and moderation system of Mastodon was criticized in interviews for the low quality of its user interface (UI), which lacked automatisation and was delegating to moderators important decisions, such as flagging of specific undesirable content, its categorisation and decisions such as temporary or permanent account suspension. The "clumsy UI" could even lead, as reported by interface administrators, to accidental account deletion. These debates within the Mastodon community brought developers to introduce in 2018-2019 an Application Programming Interface (API) that could offer better usability for instance moderators by allowing them to use third-party tools for moderation and offering a possibility to automate filtering based on custom lists of keywords that stay instance-specific.

Case study 2: Matrix.org and "protocol neutrality"

Matrix.org is a federated messaging ecosystem that proposes state-of-the-art end-to-end encryption based on the Signal protocol. The main goal of the project is to

15. *August 2018 – Read the Tea Leaves*. (2018). Read the Tea Leaves. https://nolanlawson.com/2018/08/

16. petites singularités (2021, January 15). *2021-01-23 SocialCG meeting - new fediverse users*. SocialHub. https://socialhub.activitypub.rocks/t/2021-01-23-socialcg-meeting-new-fediverse-users/1305

create an architecture able to fully tackle the interoperability problem. This interoperability is meant to become a substantial comparative advantage and enrollment factor for users (Hodgson, 2023; Hendriks, 2020; see e.g. Weinberger, 2014 for a representative specialised press article).

The challenges of moderation and debates about freedom of speech have long been discussed in connection to social media platforms, defined as arenas of public debate that have to be regulated. On the contrary, messaging applications were rarely discussed as requiring moderation and were perceived as closed environments or silos, where communications happened in private. In our study we suggest that some of the modern messaging applications can be considered as hybrids between social networks and private communication tools, because they offer a feature of public group chats or rooms (as in Matrix/Element), that can be searched and joined from the outside.

Telegram is one example of such a hybrid tool that unites features of a messenger and of a social network. However, it is centralised and as such is out of the scope of our study which only focuses on federated applications. Matrix/Element is an interesting case of such hybridity, as it offers a default public server (and therefore, a degree of centralisation), as well as a public list of rooms and channels that people can request to join. This existence of repertoires of public rooms and their potential discoverability by external users makes Matrix a semi-public space and thus introduces a demand for moderation that comes up from the user community as well as from the regulators and app markets. As our fieldwork interviews have helped to illustrate, since its beginnings, the Matrix team, unlike Mastodon's, did not take an explicitly political or ideological stance, and did not aim at providing software for specific audiences with a political agenda or engaged in political arenas, such as activists. This position, a kind of 'liberal pluralism', is reflected in the very architecture as well as the users of his system. From the point of view of the architecture, it is a federated system that bridges a great variety of different messaging tools, thus leaving a certain amount of freedom to users, allowing them to retain their usual interface, while making it possible for them to connect with others. In terms of user pluralism, Matrix has a variety of rooms addressing a wide variety of subjects, from cryptography and open-source, cryptocurrency and decentralization to psychological help, furries, subcultures and fan communities, left-wing groups and alt-right Donald Trump supporter rooms. Two of the main lingering problems for Matrix are managing spam and maintaining a decentralized reputation system -- two issues that, according to the Matrix founders, are still open for research.

During our interview with the co-founder of Matrix, Matthew Hodgson, in 2017, moderation and reputation systems had already been discussed as possible challenges for future developments. Back then, the position of Matthew Hodgson was that of radical inclusivity and free speech. In response to our question about the targeted user groups for Matrix, he said he could not be aware of all rooms and servers within Matrix since it is a federated and open source network. And even though he was aware of "some pizzagate right-wing guys using it" (cit.), he was against the idea of a master directory for all servers or of introduction of backdoors of any kind:

> "*We utterly abhor child abuse, terrorism, fascism and similar - and we did not build Matrix to enable it. However, trying to mitigate abuse with backdoors is, unfortunately, fundamentally flawed"* (Matthew Hodgson, Matrix co-founder, in a Matrix.org blog post, 2020[17])

However, in 2021, several years after our interview with Hodgson, Element, the Matrix client, was banned by the popular digital distribution service Google Play because some "abusive content" had been discovered by Google Play bots. As a consequence, moderation became an urgent issue. As a response, Matrix developed Mjolnir: a support bot for bans, redactions, anti-spam, room shutdown and other moderation activities, and a relative reputation system (published as a reputation feed) that allows anyone to produce subjective scores on users, servers, rooms or messages. The bot Mjolnir has to be set up and run on a dedicated infrastructure by the server administrators. It can ban users, delete messages and execute other moderation activities depending on the desirable configuration. One important feature of Mjolnir is to protect individual moderators from retaliation: the bans and message deletion or editing are associated to the bot's ID and not to the moderators' IDs, therefore reducing direct retaliation against moderators for their moderation actions

Our analysis shows that, in this particular socio-technical, evolving context, Matrix has opted for "protocol neutrality", i.e., not to implement any automatic moderation at the protocol level. In the above-mentioned blog post, Matthew Hodgson remarks:

17. Hodgson, M. (2020). "Combating abuse in Matrix - without backdoors". Matrix.org blog, https://matrix.org/blog/2020/10/19/combating-abuse-in-matrix-without-backdoors/

> "*The protocol's position in this solution should be one of neutrality: it should not be deciding what content is undesirable for any particular entity, and should instead be empowering those entities to make their own decisions*[18]".

Instead of baking moderation into protocols, Matrix suggests "moderation policy lists" or "ban lists" which are simple scripts stored as "room states" (configuration files with specific settings regarding content policies). These scripts can be shared across rooms and servers.

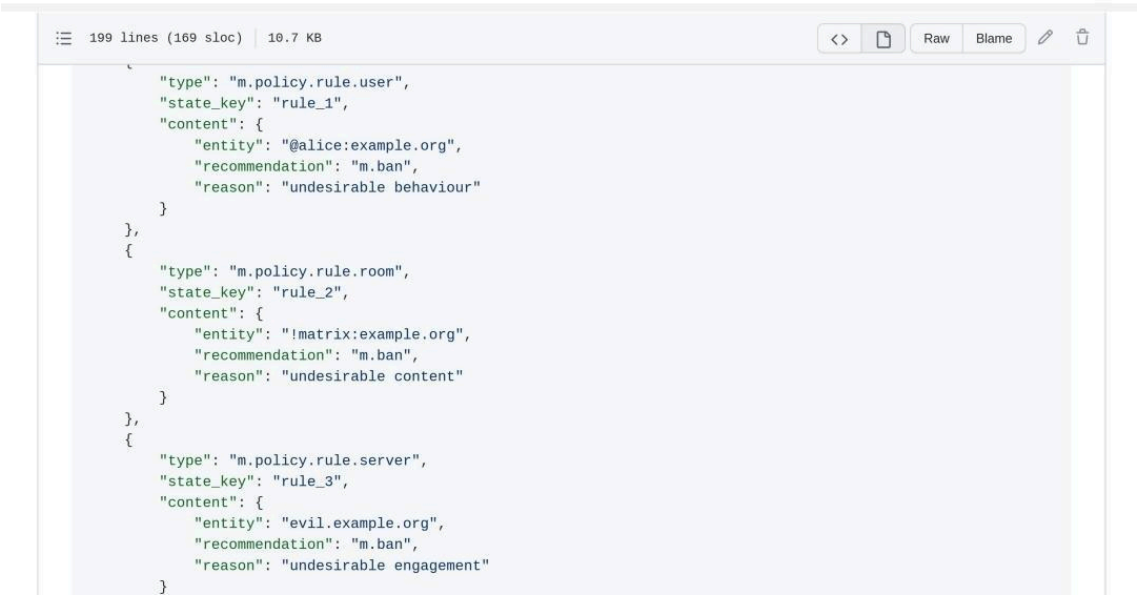

FIGURE 3: Example of a room state.

This idea of Matrix's protocol neutrality echoes well with Mastodon's attitude to machine learning-based moderation, outlined above.

The minimization of the spread of disinformation and spam appears indeed to be Matrix/Element's current main goal, to be achieved by a mix of social and technical moderation by server or instance administrators. The Matrix team urges its power-users to create their own servers rather than depend on the central default instance matrix.org, as the moderation on an independent server can be much better customised. The Mjolnir bot's capacities to assist with content moderation depend solely on the administrators of a particular server, because the bot is not centrally run by the official Matrix.org team but has to be deployed and run independently. Therefore, moderation on Matrix remains a decentralised practice that re-

18. Ibid.

lies on both community principles and technical adjustments of the existing tools. The Matrix team hopes to address this problem by also deploying a reputational system, and seeks a way for users to filter content by developing a system of open and modulable filters. As a parallel project aimed at mitigating State-based censorship, and a response to the increased risk of internet shutdowns in politically unstable regions, such as Belarus, Iran, Kirghizistan and others, Matrix has released in 2020 an alpha peer-to-peer version of its software, meant to achieve independence from Internet connections provided by telecom operators, which is however still a work-in-progress.

Case study 3: The controversy surrounding "Threads"

In May 2023, Meta launched its microblogging platform "Threads[19]" in response to what was, by far, the heaviest crisis of Twitter, following its acquisition by Elon Musk and its change of name to "X" acted soon after. In parallel, the most important migration wave from X to Mastodon in recent history unfolded in 2022-2023, followed by another wave of exodus from X to BlueSky in 2024. This underscores again the point made earlier on in the article, that social networking and microblogging platforms experience moments of great success and of decline, as technical tools and as arenas of public debate – and users' trust on and reliance upon them should not be taken for granted.

19. https://www.threads.net/; as a part of the Meta ecosystem, Threads requires an Instagram account to be operated.

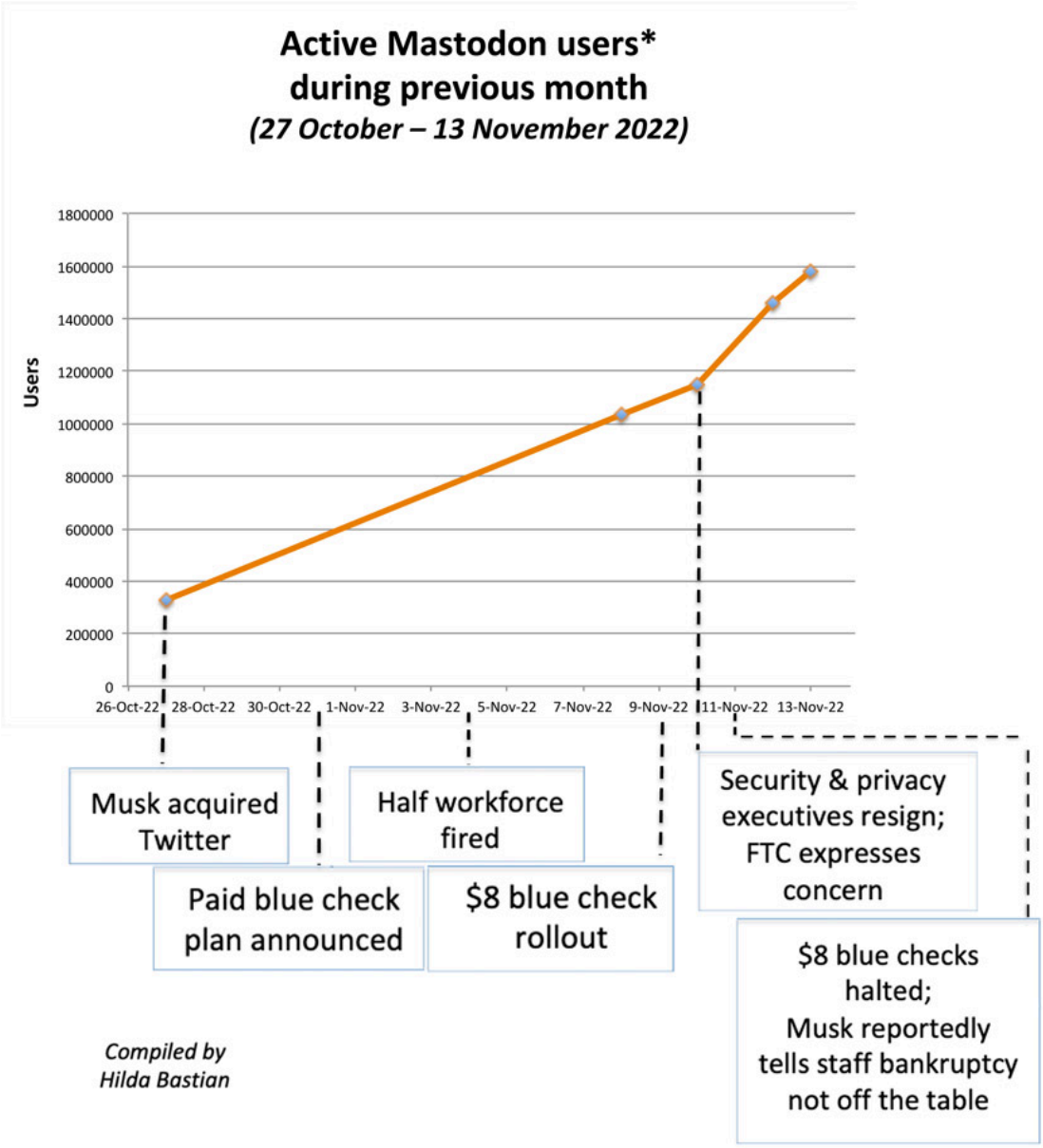

FIGURE 4: Evolution of active Mastodon users following major events in Twitter's governance after Elon Musk's acquisition. Source: Bastian, 2022.

The promise of Meta to implement the ActivityPub protocol, and rumors about background negotiations between Threads representatives and administrators of popular Fediverse instances, divided the Mastodon community.

The #fedipact movement was launched, around a dedicated webpage[20] with a manifesto and a rapidly growing list of instance administrators who signed an agreement not to federate with instances owned by Threads (as of March 2024, the list includes 682 signatures of instance administrators and the hashtag was used

20. Anti-Meta Fedipact (2025). Fedipact.online. https://fedipact.online/

by 7.49% of all active Mastodon users)[21].

The author of #fedipact initiative, user @vantablack, emphasizes that the main reason for her campaign against Threads is precisely related to its problematic approach to content moderation:

> *"At the end of the day the whole thing with The Pact blocking meta, for me at least, comes down to one fact: they are absolutely NOT going to moderate their shit properly (...) if they were any other instance they'd be defederated immediately for the shit they've aided and abbetted [sic] in and allowed to exist on their platform".*

Because of the very architecture of Mastodon and other ActivityPub-based services, federating with Threads would mean that potentially harmful content produced by Threads' users would show up in Fediverse users' feeds. Core individuals behind #fedipact assumed that Threads was a real threat to the Fediverse's unique ethics and could undermine its reputation as a "safe space".

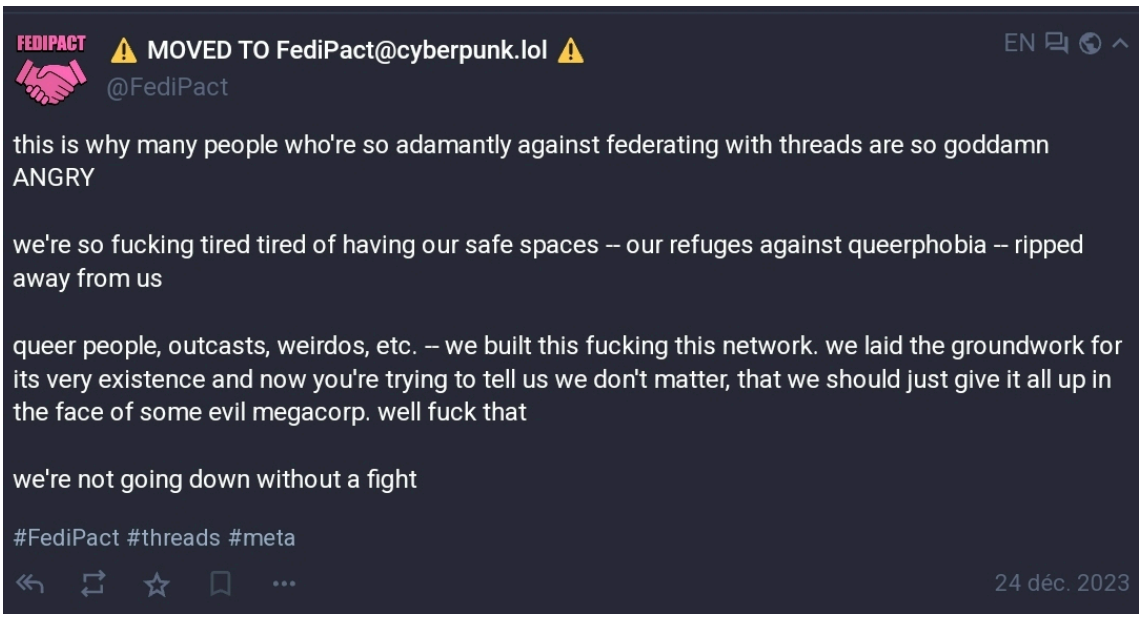

FIGURE 5: Threads is perceived as a threat by vulnerable groups. Source: @FediPact@tech.lgbt. (2025). LGBTQIA+ and Tech. *https://tech.lgbt/@FediPact/111634849265490496*

After Mark Zuckerberg's announcement on January 7, 2025 of major changes in moderation of Meta services, the biggest Mastodon instance Mastodon.social has taken a major decision to defederate from Threads.net instance and block

21. FediDB, Fediverse Network Statistics (2025). Fedidb.org. https://fedidb.org/current-events/anti-meta-fedi-pact

threads.net at the DNS level: “The new Meta policy goes against much we stand for and we will not allow to let it spill over to the Fediverse”[22].

Our analysis of the #fedipact controversy has shown that there were three main kinds of reactions by instance administrators, that we could define as techno-optimistic, protective and radical.

The first group would describe Threads’ adoption of the ActivityPub protocol as a victory of the FLOSS (Free Libre and Open Source Software) community, having a positive impact on the overall decentralization and popularization of the Fediverse. Techno-optimists would defend radical openness as opposed to the walled gardens raised by Meta’s products.

The second, and the most popular, position can be defined as protecting the federation’s alternative potential and vision, as it equates to a “de-federation” with any instance owned by Threads. As the various polls launched by Mastodon administrators show, this position has between 50 and 70% of votes[23]. The protectionists would defend a certain vision of the “Mastodon culture” and see the #fedipact as a tool of collective action against the adversarial culture and values of a proprietary platform.

Finally, the most radical group would go even further in the adoption of the #fedipact and suggested extending the “de-federation” not only to Threads, but also to Mastodon instances that agreed to federate with Threads.

The Threads controversy shows that the choice of implementing the ActivityPub Protocol has a consequence on the global governance and content moderation of the Fediverse. Choosing a federated protocol entails (potentially also negative) consequences for the federation, in terms of possibly undesirable content, posted on a platform that is considered controversial by many actors in the community, getting relayed by all instances in the federated platform. However, this moderation, in which the choice of the protocol plays a crucial role, raises an important question about the power balance between instance administrators and individual users and their right to access information. In response to this, Pixelfed (a federated equivalent of Instagram) has implemented an opt-in for users who still wish to

22. Post by @stux@mastodon.social, one of the Mastodon.social core maintainters, January 8, 2025, https://mstdn.social/@stux/113793895110300721

23. Jon. *Should the Fediverse welcome its new surveillance-capitalism overlords? Opinions differ!* The Nexus of Privacy. (December 31, 2023). https://privacy.thenexus.today/should-the-fediverse-welcome-surveillance-capitalism/#polls

see content published by Threads users[24]. This controversy also questions Mastodon's "echo-chamber"[25]effect, its dominant political culture, definition of "free speech" and tacit values.

While the ActivityPub protocol enables openness and decentralization, when it is used as a tool of collective governance and content moderation, it introduces an important risk of centralization and reshapes power structures within Mastodon and the Fediverse as a whole.

Conclusion. The "Four Cs" of federation meet content moderation

While federation is likely to pave the way for novel and potentially promising ways of content moderation, that merge aspects of community organizing, information distribution and alternative techno-social instruments, the very technical architecture that holds promise can become a weakness or a liability in particular circumstances, such as re-centralization around a small group of administrators, accident-prone interfaces, and problematic delegation chains.

In our previous research on federated architecture platforms (Ermoshina & Musiani, 2022), as well as in the present paper, we have analysed the shaping of federation in light of the 'four Cs of federation': community, customisation, compatibility and care. In the following paragraphs, we revisit these four dynamics and aspects in light of this paper's focus on content moderation.

In terms of the first C, community, (self)-governance and advancement of federated projects implies an important community-driven effort and depends on engaging a variety of service providers and clients into accepting new open protocols or new libraries, via consensus-building strategies. Our research quite clearly demonstrates the rise of a powerful and diverse community of interested actors involved in the co-production of elements (protocols, packages, libraries…) necessary to prepare the digital ecosystem for federated environments. In these environments, the community-driven effort is traceable in several aspects of the content moderation processes. First of all, the reputation of servers and rooms is collectively built, and subject to continuous evolutions; second, codes of conduct are continuously and collectively debated; third, the effectiveness of the moderation is based on the re-

24. pixelfed official account on Mastodon (@pixelfed@mastodon.social). March 22, 2024 https://mastodon.social/@pixelfed/112138026280077088

25. "the formation of groups of like-minded users framing and reinforcing a shared narrative" (Cinelli et al., 2021).

sponsiveness of instance administrators vis-à-vis the community.

The second C, customisation, highlights how federation proposes to users the option to choose among multiple service providers and migrate from one server to another without losing their social graphs. Federated architectures make it simpler to customise and localise implementations, adapting them to the needs of a specific user community without losing the ability to interact with broader networks; at the same time, implementations of a federated protocol are harder to control, and this may create security vulnerabilities across different instances or clients. In terms of moderation, this implies that moderation solutions are left on the implementation level; they do not affect the protocol itself, as summarized by the "protocol neutrality" label of Matrix.

We have identified compatibility and its challenges as the 'third C' of federation; for example, the need to implement the so-called 'backwards compatibility' that makes a harmonious transition from older to more recent protocols possible, without blocking or boycotting 'by design' some of the clients. In terms of moderation, this means that moderation solutions, as they are conducted at the implementation level, can be shared across instances, like room-states.

Finally, federation adds a layer of complexity in the governance secure messaging systems by introducing new key players, notably the system administrators, responsible for the maintenance and growth – the "care" (Denis & Pontille, 2015) – of federated infrastructures, our fourth and final 'C'. The stability of federated ecosystems depends, as well, on the successful enrollment of maintainers, that requires development of good documentation and guides with "best practices", dissemination of technical expertise through offline educational events for future sysadmins. As for moderation, the "care" aspect is made explicit by the fact that moderation solutions are implemented without harming the infrastructure and the user, and eliminating by design the possibility of backdoors.

In conclusion, in federated systems, no single entity can be counted upon for maintaining the system as a functioning one, including at the level of content moderation governance; the necessity of 'care' is distributed across the multiple sysadmins and other actors that manage the different instances in the federation. The growth of federated platforms seems to mark a turn towards community-managed 'safer spaces', with more power delegated to human moderators. However, we should keep in mind that this introduces new risks of the re-centralization of power within federated networks, requiring more research on the role of infrastructure maintainers, administrators and moderators, besides the core-set of protocol de-

signers – a research agenda that this paper has started to unfold. Federated messengers have many challenges, including spam, reputation system, as well as discoverability of contacts and content that becomes harder without a centralized registry; however, they are seen as a promising alternative by those users we have called 'disinformation refugees' (Ermoshina & Musiani, 2022) -- users who abandon currently dominant platforms due to their disillusionment about disinformation or hate speech.

References

Albrecht, M. R., Mareková, L., Paterson, K. G., & Stepanovs, I. (2022). Four Attacks and a Proof for Telegram. *2022 IEEE Symposium on Security and Privacy (SP)*, 87–106. https://doi.org/10.1109/SP46214.2022.9833666

Bastian, H. (2022). *Mastodon growth numbers might not mean what you think they mean*. Absolutely Maybe – Plos Blogs. https://absolutelymaybe.plos.org/2022/12/05/mastodon-growth-numbers-might-not-mean-what-you-think-they-mean/

Bennett, E. A., Cordner, A., Klein, P. T., Savell, S., & Baiocchi, G. (2013). Disavowing politics: Civic engagement in an era of political skepticism. *American Journal of Sociology*, *119*(2), 518–548. https://doi.org/10.1086/674006

Blondiaux, L. (2017). *Le nouvel esprit de la démocratie: Actualité de la démocratie participative*. Média Diffusion. https://journals.openedition.org/lectures/2032

Bono, C. A., La Cava, L., Luceri, L., & Pierri, F. (2024). An exploration of decentralized moderation on Mastodon. *ACM Web Science Conference*, 53–58. https://doi.org/10.1145/3614419.3644016

Butcher, M. (2014). *End of an era as VKontakte founder Durov sells his stake to Russian mobile giant*. TechCrunch. https://techcrunch.com/2014/01/24/end-of-an-era-as-vkontakte-founder-durov-sells-his-stake-to-russian-mobile-giant/

Casilli, A. (2014). *Quatre thèses sur la surveillance numérique de masse et la négociation de la vie privée [Four theses on digital mass surveillance and the negotiation of privacy]* (Rapport du Conseil d'Etat, pp. 423–434). https://shs.hal.science/halshs-01055503

Cavoukian, A. (2012). Privacy by design [leading edge]. *IEEE Technology and Society Magazine*, *31*(4), 18–19. https://doi.org/10.1109/MTS.2012.2225459

Cinelli, M., Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9). https://doi.org/10.1073/pnas.2023301118

Davisson, A., & Alati, K. (2024). "Difficult to just exist": Social media platform community guidelines and the free speech rights of sex workers. *Social Media + Society*, *10*(1), 20563051231224270. https://doi.org/10.1177/20563051231224270

Deller, R. A. (2019). Safer spaces. In *Routledge handbook of radical politics* (pp. 222–239). Routledge.

Denis, J., & Pontille, D. (2015). Material ordering and the care of things. *Science, Technology, &*

Human Values, *40*(3), 338–367. https://doi.org/10.1177/0162243914553129

Ermoshina, K., Halpin, H., & Musiani, F. (2017). Can Johnny build a protocol? Co-ordinating developer and user intentions for privacy-enhanced secure messaging protocols. *Proceedings 2nd European Workshop on Usable Security*. European workshop on usable security, Paris, France. https://doi.org/10.14722/eurousec.2017.23016

Ermoshina, K., & Musiani, F. (2019). Hiding from whom?: Threat models and in-the-making encryption technologies. *Intermédialités*, *32*. https://doi.org/10.7202/1058473ar

Ermoshina, K., & Musiani, F. (2021). The Telegram ban: How censorship "made in Russia" faces a global Internet. *First Monday*, *26*(5).

Ermoshina, K., & Musiani, F. (2022). *Concealing for freedom*. Mattering Press. https://doi.org/10.28938/9781912729227

Fuller, M. (Ed.). (2008). *Software studies: A lexicon*. The MIT Press. https://doi.org/10.7551/mitpress/9780262062749.001.0001

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

Graber, J. (2022). *Designing Decentralized Moderation"*. Medium. https://jaygraber.medium.com/designing-decentralized-moderation-a76430a8eab

Guélou, A. (forthcoming). *Configurations techno-éthiques pour les médias sociaux décentralisés et fédérés [Techno-ethical configurations for decentralised and federated social media]* [Ongoing PhD dissertation, Technical University of Compiègne]. https://metacartes.net/numerique-ethique/?LaFeditheseTheseDeDoctoratIntituleeConf

Hassan, A. I., Raman, A., Castro, I., Zia, H. B., De Cristofaro, E., Sastry, N., & Tyson, G. (2021). *Exploring content moderation in the decentralised web: The Pleroma case*. https://doi.org/10.48550/ARXIV.2110.13500

Hendriks, F. (2020). *Analysis of key management in Matrix* [Doctoral dissertation, Radboud University]. https://www.cs.ru.nl/bachelors-theses/2020/Floris_Hendriks___4749294___Analysis_of_key_management_in_Matrix.pdf

Hodgson, M. (2023). *Designing matrix: A global decentralised end-to-end encrypted communication network*. https://www.usenix.org/conference/srecon23emea/presentation/hodgson

Kwet, M. (2020). Fixing social media: Toward a democratic digital commons. *Markets, Globalization & Development Review*, *5*(1). https://doi.org/10.23860/MGDR-2020-05-01-04

Lawson, N. (2018). Mastodon and the challenges of abuse in a federated system. *Read the Tea Leaves. Software and Other Dark Arts*. https://nolanlawson.com/2018/08/

Musiani, F. (2013). Dangerous liaisons? Governments, companies and internet governance. *Internet Policy Review*, *2*(1). https://doi.org/10.14763/2013.1.108

Myers West, S. (2018). Cryptographic imaginaries and the networked public. *Internet Policy Review*, *7*(2). https://doi.org/10.14763/2018.2.792

Rosanvallon, P., & Goldhammer, A. (2008). *Counter-democracy: Politics in an age of distrust* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511755835

Rozenshtein, A. Z. (2022). Moderating the fediverse: Content moderation on distributed social media. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4213674

Snowden, E. (2019). *Permanent Record*. Henry Holt and Company.

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326

Tucker, C. (2019). Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. *Review of Industrial Organization*, *54*(4), 683–694. https://doi.org/10.1007/s11151-019-09693-7

Weinberger, M. (2014). *Matrix wants to smash the walled gardens of messaging*. Computerworld. https://www.computerworld.com/article/1384128/matrix-wants-to-smash-the-walled-gardens-of-messaging.html

Zuckerman, E. (2010). Intermediary censorship. In R. Deibert, J. Palfrey, R. Rohozinski, & J. L. Zittrain (Eds.), *Access Controlled* (pp. 71–86). The MIT Press. https://doi.org/10.7551/mitpress/8551.003.0010

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

UOC R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies