Safer spaces by design? Federated architectures and alternative socio-technical models for content moderation

Ksenia Ermoshina, Francesca Musiani

Safer spaces by design?
Federated architectures and alternative socio-technical models for content moderation

Ksenia Ermoshina
ksenia.ermoshina@cnrs.fr

Francesca Musiani
francesca.musiani@cnrs.fr

Centre Internet et Société
Centre national de la recherche scientifique (CNRS), Paris, FR

Paper presented to the Annual Symposium of the Global Internet Governance Academic Network (GigaNet), 2022 (virtual)

Draft/work in progress. Please do not cite or quote without the authors' permission.

Abstract. Users of secure messaging tools, especially in communities attuned to the risks of State-based and other forms of censorship, are becoming increasingly skeptical about the fact of delegating their data to centralized platforms, endowed with substantial power to filter content and block user profiles. This paper analyses the role that informational architectures and infrastructures in federated social media platforms play in content moderation processes. Alongside privacy by design, the paper asks, is it possible to speak of online "safe(r) spaces by design"? The paper argues that federation can pave the way for novel practices in content moderation governance, merging community organizing, 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 need to be examined to provide a complete picture of the potential of federated models.

Keywords. Federation; content moderation; censorship; secure messaging; governance

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 catalyzed 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 centralized and proprietary messengers and social media platforms (Ermoshina & Musiani, 2022), and the need to look for alternatives, both socio-technical and political. This echoes with the well-documented mistrust towards representative democracies and critique of traditional forms of political participation (Rosanvallon & Goldhammer, 2008; Blondiaux, 2017; Bennett et al., 2013). 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 p2p apps 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 alleviate the very high degree of personal responsibility held by a centralized 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,

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

favoring 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 those users is legitimate or not, our paper focuses on the role of informational architectures and infrastructures of federated social media platforms in content moderation processes. Alongside privacy by design (see Cavoukian, 2012), can we speak of online "safe spaces by design" ?

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 Twitter to decentralized and open source tools that constitute the Fediverse, where Mastodon is an outstanding example. 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).

We argue that federation can pave the way for novel practices in content moderation governance (Hassan, 2021), merging community organizing, information distribution and alternative techno-social instruments to deal with online harassment, hate speech or disinformation, proposing 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 with a Russian feminist activist, L., who critically assessed the techno-optimist promise of absolute safety and 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. This attitude towards online communication tools is well described by her drawing, depicting potential threats as alligators.

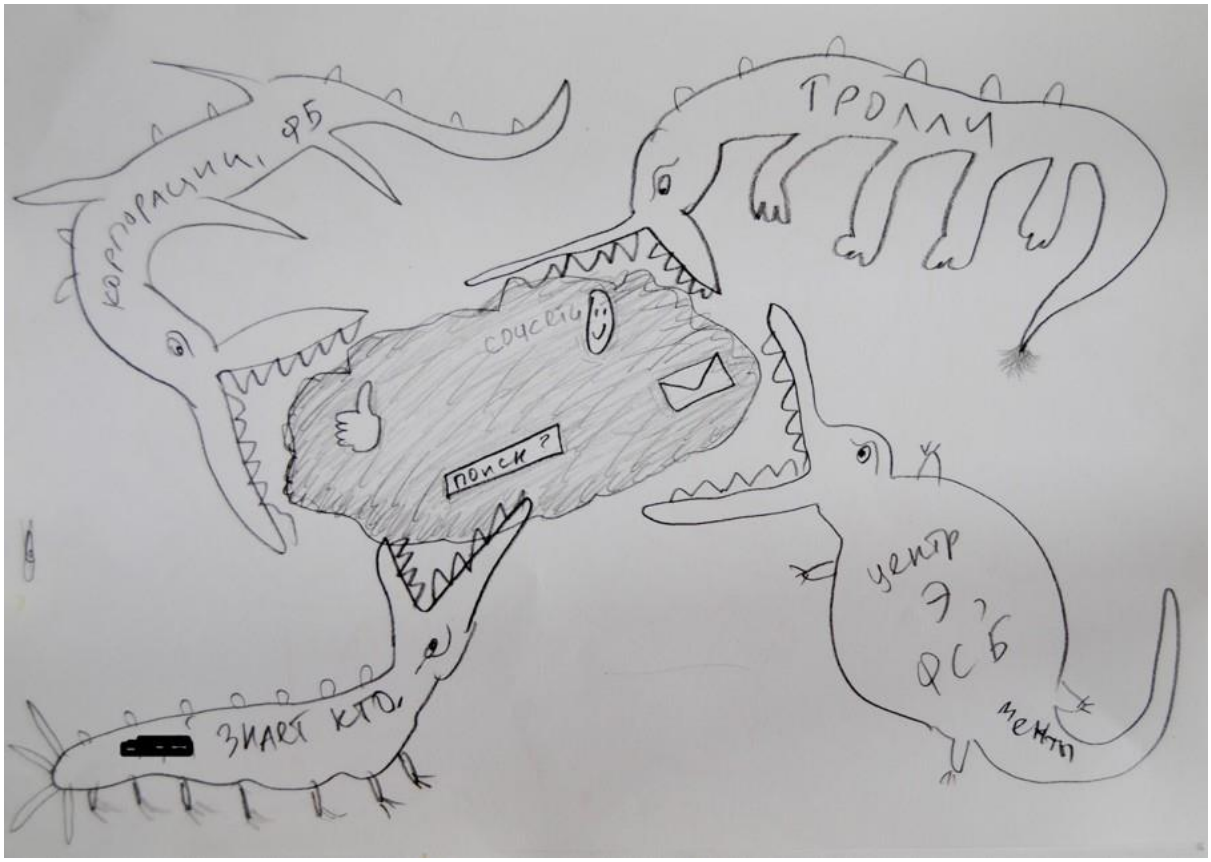

Image 1. "How I see threats on the Internet" (collected during an interview in Russia, as part of the fieldwork on users of secure messaging apps, for the NEXTLEAP project)

This paper analyses the Fediverse as an alternative model for content distribution and moderation, describing briefly its founding principles and key projects. We pay particular attention to their interfaces and the underlying protocols and architectures of these tools (for example, the core role played by ActivityPub protocol and the interoperability it offers). Understanding information architectures from an STS perspective (Star, 1999 ; Fuller, 2008), we analyze 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 centralized platforms such as Twitter 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. The empirical part of the paper is organized around two case studies, Mastodon and Matrix.org.

Methodology

While academic researchers finally start to examine the questions of content moderation and governance of the Fediverse (see for example Rozenshtein, forthcoming 2023), our paper proposes an STS approach with a particular attention to the architectural and infrastructural aspects of federated platforms. This paper is a work-in-progress based on an ethnographic study

of Mastodon, Matrix and Delta Chat communities, including interviews with users and developers of federated messaging applications and the Fediverse server or instance administrators, 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[2]. This research was initially conducted in the frame of the NEXTLEAP project (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).

Digital migrations and the rise of the Fediverse

In our study conducted between 2016 and 2018 with 90+ users of end-to-end encrypted messaging apps, we explored, besides other research questions, the motivations behind user preferences for a particular messenger. In the context of a vibrating market of "privacy by design" apps, why do users trust one tool more than the other? Interestingly enough, for the majority, 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. For instance, the success of Telegram in Russia, that we thoroughly analyzed in a dedicated paper (Ermoshina & Musiani, 2021); had very little to do with the quality of the actual cryptographic protocols used by Telegram, that were largely criticized by the security community. 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 in the last six years. 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 criticized 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. 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, Pavel Durov sold the 'made in Russia' social network Vkontakte or Vk.com to the Russian oligarch Usmanov, and as a consequence, 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.

[2] https://www.w3.org/community/socialcg/

However, digital migration is not a linear process, and the metaphor of migration itself has its limits. Unlike geographical migration, a digital one 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.

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, forthcoming 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 the users' needs for plurality of tools and content forms.

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. Another reason expressed by one of our interviewees, an Austrian Mastodon instance[3] administrator, *"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"*. 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 for 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.

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

Unlike the centralized Twitter, Mastodon is based on a federated architecture that is built on what we called the "four C's": community, compatibility, customization and care (Ermoshina & Musiani, 2022).

The community and care aspects are reflected in the way Mastodon ecosystem is regulating itself, where 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 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 behavior of the users of a particular instance, its values, fields of interest, acceptable and unacceptable content.

Hailed by some as the "nazi-free Twitter"[4], 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[5] 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[6] that he did not have any control of the situation because of the federated nature of Mastodon and the openness of its source code. The Mastodon community, however, found a way to react to this misusage of their platform, embedded in the very architecture of Mastodon: the right-wing instances were simply blocked by many other instances, therefore isolated or "unfederated".

Mastodon's federated architecture actually offers users a different experience as compared to 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" published by users of their local instance, but also other instances that their instance is "federating" "Unfederation" is comparable to "unfollowing" but on the level of a server and is usually a decision taken by an instance administrator together with its user community.

[4] https://www.esquire.com/lifestyle/a22777589/what-is-mastodon-twitter-platform/
[5] https://blog.joinmastodon.org/2019/07/statement-on-gabs-fork-of-mastodon/
[6] https://www.theverge.com/2019/7/12/20691957/mastodon-decentralized-social-network-gab-migration-fediverse-app-blocking

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 7500 moderators for 2 billion users, in Mastodon it could be 1 to 500 on some instances, but 1 to 5000 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)

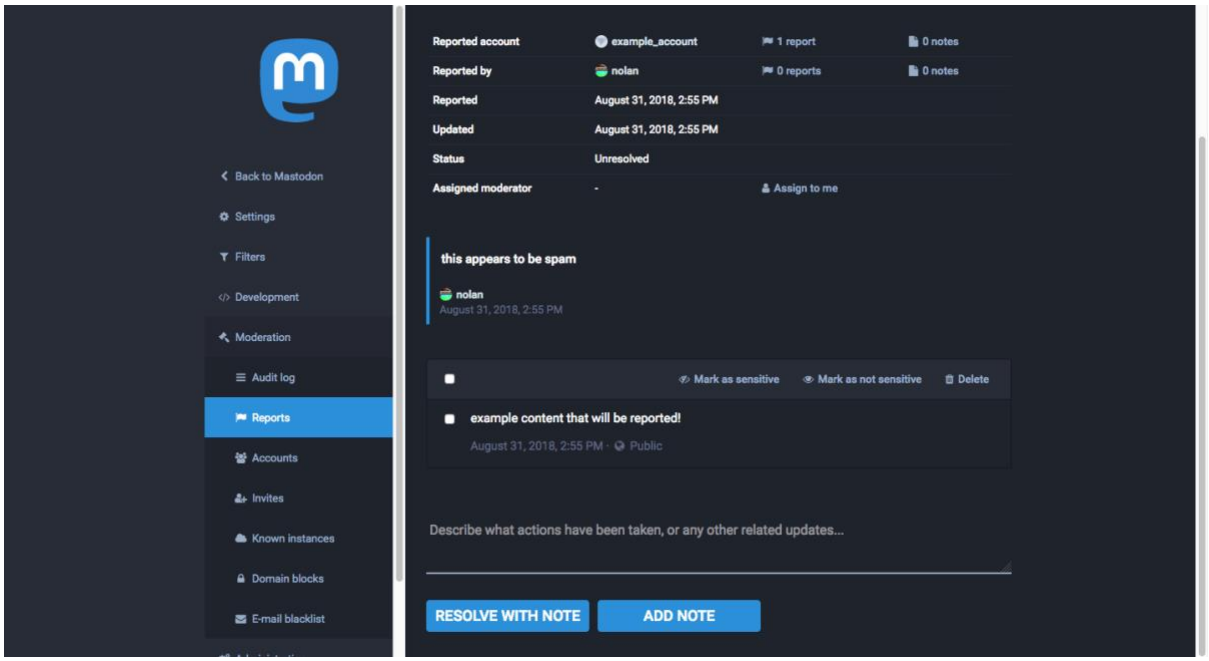

Image 2. The moderator interface for handling reports in a Mastodon instance

One of the attempts to automate moderation is 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" (Graber, 2021). 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 have 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 centralization; but the implications of this vis-à-vis machine learning is that the learning datasets risk to be rather small and might make little to no sense, thus reducing the value and validity of these approaches.

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 of our interviewees still think this very "qualitative" and "human" approach is ultimately the best tool to moderate an instance.

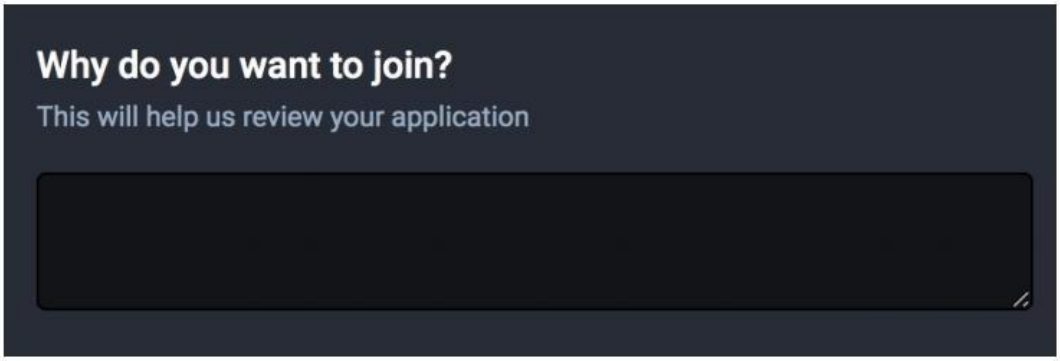

Image 3. Standard question asked to a user wishing to join a Mastodon instance.

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[7]. Our interviews with instance moderators and active users, as well as desk research mapping debates on content moderation in Mastodon have enabled us to analyze actual content moderation practices and the 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,

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

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[8].

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[9] 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 automatization and was delegating to moderators important decisions such as flagging of specific undesirable content, its categorization 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 2019 an Application Programming Interface (API) that could offer better usability for instance moderators by allowing them to use third-party tools for moderation.

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, as underlined in several articles[10], is to 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. Since its beginnings, the Matrix team 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 ahrenas, 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

[8] https://nolanlawson.com/2018/08/
[9] https://socialhub.activitypub.rocks/t/2021-01-23-socialcg-meeting-new-fediverse-users/1305
[10] E.g. https://www.computerworld.com/article/2694500/matrix-wants-to-smash-the-walled-gardens-of-messaging.html

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, and need to be supported by a 'morally neutral' positioning.

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 a 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"* (interview, Matthew Hodgson, Matrix co-founder)

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.

Matrix has opted for "protocol neutrality", i.e., not to implement any automatic moderation at the protocol level:

> "*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*" (interview, Matthew Hodgson)

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.

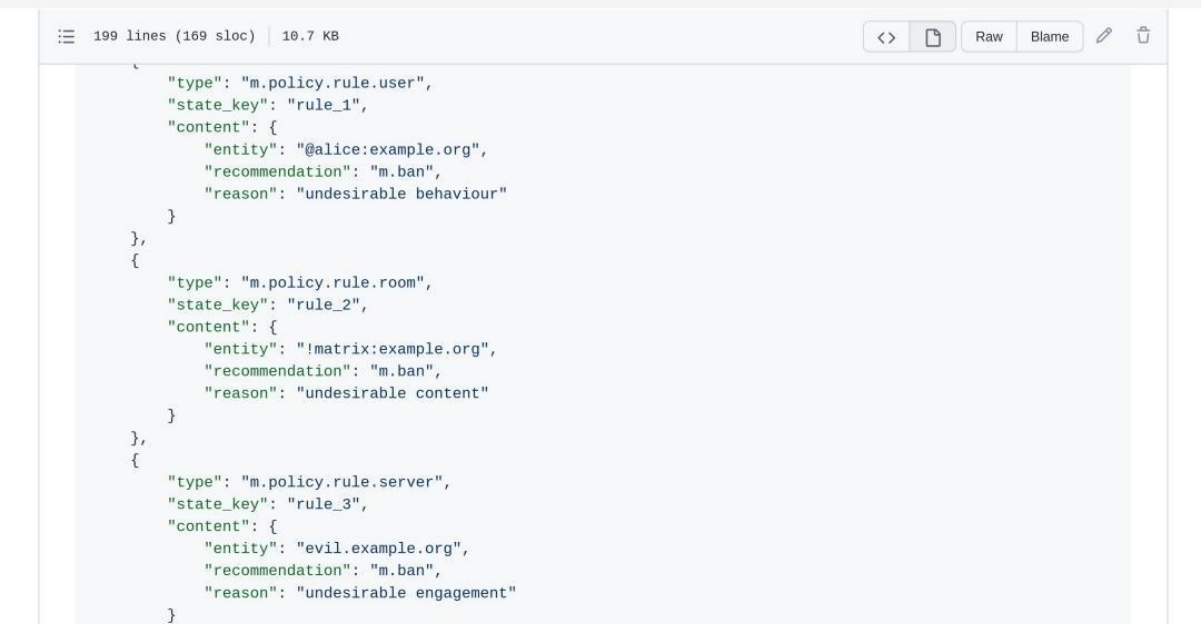

Image 4. Example of a room state

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 hopes to address this problem by 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.

Conclusion. **Revisiting the "Four C's" of federation**

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 and Musiani, 2022), as well as in the present paper, we have analysed the shaping of federation as both an infrastructural and a social experiment. We have seen how, in different projects, developers seek to achieve a compromise between high levels of security and better usability, in a constant dialogue with 'ideological' motivations such distributing responsibilities onto a larger number

of actors, and proposing particular definitions of online freedom, such as giving users the choice of the level of autonomy they wish to achieve. We have previously systematized the results of our research as the 'four C's of federation': community, compatibility, customization and care. We revisit these four aspects below, paying particular attention to this paper's focus on 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 a 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, likened to a "living thing"; second, codes of conduct are continuously and collectively debated ; third, the effectiveness of the moderation is based on the responsiveness of instance administrators vis-à-vis the community.

The second C, customization, 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 customize and localize 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 'safe 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 designers – 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.

Bibliography

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.

Blondiaux, L. (2017). Le nouvel esprit de la démocratie: actualité de la démocratie participative. Média Diffusion.

Casilli, A. (2015). « Quatre thèses sur la surveillance numérique de masse et la négociation de la vie privée », in *Rapport du Conseil d'Etat*, pp. 423-434.

Cavoukian, A. (2012). Privacy by design [leading edge]. *IEEE Technology and Society Magazine*, *31*(4), 18-19.

Denis, J., & Pontille, D. (2015). Material ordering and the care of things. *Science, Technology, & Human Values*, *40*(3), 338-367.

Ermoshina, K., & Musiani, F. (2018). Hiding from whom? Threat models and in-the-Making encryption technologies. Intermédialités: histoire et théorie des arts, des lettres et des techniques/Intermediality: History and Theory of the Arts, Literature and Technologies, (32).

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

Ermoshina, K., Halpin, H., & Musiani, F. (2017). Can Johnny build a protocol? co-ordinating developer and user intentions for privacy-enhanced secure messaging protocols. In European Workshop on Usable Security.

Fuller, M. (Ed.). (2008). Software studies: A Lexicon. Cambridge, MA: The MIT Press.

Graber, J. "Designing Decentralized Moderation" published on Medium on January 21, 2022; https://jaygraber.medium.com/designing-decentralized-moderation-a76430a8eab

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. arXiv preprint arXiv:2110.13500.

Kwet, M. (2020). Fixing Social Media: Toward a Democratic Digital Commons. Markets, Globalization & Development Review, 5(1).

Lawson, N. (2018) Mastodon and the challenges of abuse in a federated system; https://nolanlawson.com/2018/08/

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

Rosanvallon, P., & Goldhammer, A. (2008). Counter-democracy: Politics in an age of distrust (Vol. 7). Cambridge University Press.

Rozenshtein, A. Z. (forthcoming 2023). Moderating the Fediverse: Content Moderation on Distributed Social Media. *Journal of Free Speech Law, 2* (available as pre-print on SSRN at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213674)

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

Star, S. L. (1999). The ethnography of infrastructure. American Behavioral Scientist, 43(3), 377-391.

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.

Zuckerman, E. (2010). “Intermediary Censorship”, in Deibert, R., Palfrey, J., Rohozinski, R., & Zittrain, J. (eds.) *Access controlled: The shaping of power, rights, and rule in cyberspace*. Cambridge, MA: the MIT Press, pp. 71-86.