Findings Report: Governance on Fediverse Microblogging Servers

Erin Kissane and Darius Kazemi with the support of the Digital Infrastructure Insights Fund
Correspondence: fediversalistpapers@gmail.com
August 20, 2024
[PDF version / web version]
[Quick-Start Guide to Fediverse Governance Decisions]
[Fediverse Governance Opportunities for Funders & Developers]

Contents of this report

How to Use These Findings **4**
Suggested reading pathways 4
Project Introduction **6**
Our goals 7
The stakes 8
The governance knowledge gap 8
Realistic risks and mitigations 9
What we mean by the Fediverse 10
What we mean by governance 10
Methods 12
Why we focused on medium-sized Mastodon and Hometown servers 12
How we chose our interviewees 13
What we missed 15
How we assembled this report 15
Brief glossary 16
Special thanks 17
Section One: Overall Observations **18**
1. The big themes 18
2. The risks 26
Class 1: Risks framed by our participants as manageable 26
Class 2: Sources of moderate unresolved frustration or anxiety 27
Class 3: Sources of broader and more intense anxiety 28
3. High-level recommendations 30
Best practices for server teams 30
Opportunities for addressing unmet needs and unmitigated risks 31
Section Two: Moderation **34**

Introduction 34
Key observations 34
The anatomy of Fediverse moderation 36
1. Registration 37
1.1 Registration by application 37
1.2 Closed and invite-only registration 39
1.3 Open registration 40
2. Rules and guidelines 42
2.1 Documentation types and links 42
2.2 Rule-making as moderation 43
2.3 Initial rule-making process as the first step toward governance 44
3. Moderation basics 46
3.1 Everyday moderation tasks 46
3.2 Variations across topics and by individual user 47
4. Complex moderation actions and decisions 49
4.1 Vibes and norms 49
4.2 Collaborative decision-making 52
4.3 CSAM and copyright complaints 54
4.4 Moderator mental health 54
4.5 Proactive work to reduce moderation load 55
5. Moderation teams 57
5.1 Finding the right people 57
5.2 Onboarding and training 60
6. Additional moderation resources 61
Section Three: Server Leadership 62
Introduction 62
Key observations 63
1. Three models of server governance 65
1.1 Independent top-down governance 65
1.2 Cooperative governance 67
1.3 Non-profit entities as a middle path 69
2. Specific structures and patterns 72
2.1 Membership discussions and meetings 72
2.2 Boards 72
2.3 Working groups 72
2.4 User advisory groups 72
3. Paths for exploration 74
3.1 Connecting people-people and tech people 74
3.2 Easing institutions into the Fediverse 75

3.3 Making pathways to greater participation 77
4. Governance Resources 78
Section Four: Federated Diplomacy 79
Introduction 79
Key observations 79
The diplomatic layer of governance is largely undocumented 80
Policy and structure vs. technical tools 80
1. Federation as remote moderation 81
2. Whether and when to limit and silence other servers 82
3. Threads federation as a governance stress test 85
4. The potential of federation policies 87
Section Five: Tooling 89
Introduction 89
Key observations 89
1. Moderation tools 90
1.1 Documentation and onboarding 90
1.2 Dealing with volume 90
1.3 Lack of context 93
1.4 Collaboration between local moderators 94
1.5 Communication between moderator and user 94
1.6 Shared blocklists and shared blocks 96
1.7 Account registration control 97
1.8 User-facing generic moderation account 98
1.9 Internal moderation team communication 99
1.10 Content filtering 99
2. Different forms of federation 100
3. Identity and data transfer 101
4. Conceptual location of tools 102
5. Financial tools 103
5.1 Cooperative decision-making 103
5.2 Self-limits on financial capacity 104
5.3 A gap in fiscal sponsorship 104
6. Legal compliance tools 105
7. Federation of moderation decisions 105
8. Inter-server admin comms 108
9. Tooling recommendations 109
Section Six: The Case for the Fediverse 112

Project Introduction

We proposed this project in the fall of 2023 based on our shared sense that the Fediverse's history of resilience and expansion positions it as one of our best chances to allow more people to **maintain strong social connections online while escaping the behavioral manipulation, pervasive surveillance, and capricious governance** that characterizes large-scale centralized social platforms.

Initial research question: "What are the most effective governance and administration models/structures in place on medium-to-large sized Fediverse servers, and what infrastructural gaps (human and digital) persist?"

Our rationale at the project's outset: "The Fediverse's rapid expansion brings both opportunities and multifaceted risks. Our research seeks to identify current server administrators' most promising models for mitigating those risks and outline the biggest and most important gaps in risk mitigation, with the aim of helping the broader Fediverse level up governance quickly, safely, and collaboratively."

We were drawn to this research question because the socio-technical aspects of Fediverse governance often seem opaque from the outside—from outside any given server, and especially from outside the Fediverse. Most servers offer some documentation about their practices and a few offer extensive explanations and policies, but whole swathes of knowledge about the aspects of server management that extends beyond the more purely technical concerns of hosting, provisioning, and technical upkeep exists only as *insider knowledge*.

Above all, we wanted to understand more about what happens behind the curtain of Fediverse server operation, and distribute this knowledge widely to help other server teams level up together—and perhaps to uncover characteristics of server governance that might be meaningful to others trying to build sustainable alternatives to centralized commercial platforms, whether on the Fediverse or elsewhere.

Having completed our initial inquiry, we're optimistic that:

- thoughtfully governed, medium-sized Fediverse servers are especially well positioned to offer a model of high-context, culturally sensitive online community that outperforms most interactions with centralized platform governance;
- the Fediverse's combined emphasis on the sovereignty of local norms and a federated form of network diplomacy *can* offer a real and optimistic challenge to the dead end of centralized content moderation at scale; and
- the emergent processes and technologies of the Fediverse *can* form a part of what media researcher (and Fediverse server operator) Nathan Schneider calls the "governable stack,"

which he defines as "webs of tools and techniques that can support self-governing online communities."

But, crucially, **we don't think that the Fediverse is likely to realize these potential benefits without ongoing and intentional emphasis on—and funding for—addressing the cultural, financial, legal, and technical governance needs and gaps** highlighted by our research participants.

How to Use These Findings

To make our research as useful as possible for multiple audiences, we've organized our findings and recommendations into three documents:

- The document you're reading now, our **Findings Report** (~40,000 words), is the most comprehensive record of our observations and recommendations. It's divided into six sections, and opens with a discussion of the project's stakes, goals, terms, methods, and risks, which we encourage anyone who wants to engage with the findings to read to get a sense of what we're trying—and not trying—to accomplish here, and why. After this:
    - **Section One** lays out our overall observations, the kinds of risks articulated by our participants, and our most opinionated recommendations for addressing these risks.
    - **Sections Two through Five** walk through our observations about four different facets of governance on the Fediverse (**Moderation**, **Server Leadership**, **Federated Diplomacy**, and **Tooling**).
    - **Section Six** includes a collection of the most hopeful and enthusiastic comments our participants made about their experiences with and hopes for the Fediverse; these passages were too heartening to leave out.
- The second document, the **Quick-Start Guide to Fediverse Governance Decisions** (~2,000 words) is an abbreviated introduction and a densely hyperlinked alternate path into the full Findings Report *for people who run or are considering running a Fediverse microblogging server*.
- The third document, **Fediverse Governance Opportunities for Funders & Developers** (~4,000 words), is a condensed version of our findings *for individuals and institutions interested in building and supporting stronger infrastructure for Fediverse governance*, also with links to more comprehensive information in the full Findings Report.

Suggested reading pathways

If you're **a relatively new Fediverse administrator, moderator, or other participant in governance** interested in applying our participants' insights to your own work, the simplest way into these

findings will probably be to read the **Quick-Start Guide** and the linked sections of the full **Findings Report** that draw your attention.

If you're a **part of the philanthropic ecosystem** interested in the ways of strengthening Fediverse governance that emerged from our research, we recommend starting with the **Opportunities for Funders and Developers** document, and the **introductory material** and **Sections One and Six** of the present document.

If you're **a developer (or representing a group of developers)** and want to contribute meaningfully to Fediverse governance problems we identify in our research, we recommend reading the **Opportunities for Funders and Developers** document, the **introductory material** and **Sections One**, **Five**, and **Six** of the present document.

If you're already **deeply engaged with the Fediverse, but not prepared to read 40,000 words of material in order**, we suggest beginning by reading our **introductory material** and **Section One** in the present document and scanning the **Observations** subsections at the beginning of **Sections Two through Five** to determine which other sections will be useful to you. (**Section Six** is a great chaser!

Our goals

We—Erin and Darius—are both participants in the Fediverse and in conversations about the Fediverse: Darius as a server operator, maintainer of the Hometown Mastodon fork, and advocate of independent federated social media; Erin as an internet community person and Fediverse member engaged in trying to make sense of socio-technical patterns and norms on decentralized social media systems.

In the simplest terms, we're trying to establish **how governance happens** across participating servers and teams. This report and its accompanying documents represent our attempt to understand and document existing governance systems, practices, concerns, and aspirations across a sample of thoughtfully but differently governed Fediverse microblogging servers.

Our short-term goal is to help **disseminate the substantial body of governance expertise that already exists within server teams** on the Fediverse in the hope of easing the burden on small and medium-sized server teams and helping more teams develop more sustainable practices.

In the longer term, we hope that our work here will:

- promote the funding and development of better governance tooling,
- lay the foundation for systems that can guide new and potential Fediverse members to servers that meet their governance needs,

- enhance the overall resilience of the network, and
- ultimately make the Fediverse—and perhaps other networks!—a better place for more people.

The teams we spoke with were keenly aware of the necessary trade-offs of governance, and no two server teams described their governance responsibilities, aspirations, and anxieties in the same way. Our findings will proceed in the way that seems most true to the expertise and insights our participants generously shared: multi-voiced, grounded in specifics, and open to many paths.

The stakes

The "Fediverse" network of largely open-source, interoperable internet services expanded from approximately 6,000 known servers in early 2022 to more than 29,000 in the summer of 2024, and from about 2M user accounts to more than 10M over the same period, according to FediDB. Not all Fediverse services can reasonably be classified as "social media" networks, but the vast majority of Fediverse accounts are on ActivityPub-based social media services—and the majority of these are on the Mastodon microblogging service.

The Fediverse's expansion—and the arrival of other decentralized social networks including Bluesky and Meta's Threads platform—takes place at a moment of reckoning for the centralized social media platforms that have dominated the past roughly fifteen years of online sociability. Although these platforms remain globally dominant, their growth has slowed or stagnated in the US and Western Europe while regulatory attention to the governance of centralized platforms has intensified. In the microblogging sector, the decline is particularly stark: usage of Twitter/X dropped by an estimated 30% in the US between 2023 and 2024.

Although the Fediverse is still very small in comparison to the largest centralized platforms, we think its growth since 2022 suggests that it presents a viable *technical* alternative to the big centralized platforms. But **whether the Fediverse can realize its potential as a home for better and healthier individual experiences—and a fertile ground for community experiments—depends as much on its systems of governance as on its user-facing features and interfaces**.

So: **Can the Fediverse's systems of governance ultimately outperform those of centralized tech companies?** We proposed this project because we think a crucial step in finding the answer to that question is to establish **what those systems of governance actually are** and **how they work today**.

Hundreds of interview-transcript pages later, we think it's clear that the servers we studied offer real-world examples of governance that differ from centralized platforms **not only in scale, but in kind**—and that despite the network's complexity and persistent opacity, **many of the the structural possibilities the Fediverse offers allow for the flourishing of better and more humane ways of managing human interactions online**.

The governance knowledge gap

Mastodon and many other Fediverse services present themselves as more ethical alternatives to centralized corporate platforms' opaque algorithms, ad-centric business models, and capricious leadership. The Mastodon project's primary user-facing website, JoinMastodon.org, positions Mastodon as "Social networking that's not for sale" and argues that "Your home feed should be filled with what matters to you most, not what a corporation thinks you should see."

When it comes time to actually create an account, however, potential Mastodon members encounter a gap in information that will be crucial to their experience: they can choose to create an account on the project's flagship server, or to choose an alternate server featured on the JoinMastodon.org site, but beyond reference to the existence of a baseline "Mastodon Covenant," they receive no guidance about how any of the servers they can choose are governed, nor about how those servers' governance practices will affect their experience of the Fediverse, nor about which kinds of factors they should be evaluating when they make their choice of server.

This opacity about governance isn't an oversight in Mastodon's pitch to potential members: It reflects a real gap in the Fediverse's understanding of its own governance systems and practices. **No repository of structured (or unstructured) data about the way each Fediverse server is governed currently exists, and the network's emerging governance systems, processes, and norms are largely informal and undocumented beyond the existence of rules and codes of conduct for individual servers.** As a result, people who want to join a well-governed server are left in the dark about how to evaluate their choices, while people who want to *run* well-governed servers have few resources to help them understand the problem space and assemble appropriate and effective governance systems and processes.

Realistic risks and mitigations

We also think the work of documenting existing governance processes is essential—and urgent—right now, because the Fediverse's expansion brings risks as well as opportunities. Getting more people involved in free and open networks can be a social good, but the Fediverse has historically had difficulty maintaining instance-level stability—instances implode, often due to overwork and under-funding, but also to governance problems. Additionally, we believe that the transition to truly mass scale is likely to test the ecosystem's ability to handle the big content-based threats facing commercial networks, including CSAM, spam, coordinated covert influence campaigns, and hateful and violence-inciting speech.

In Scaling Trust on the Web, a major trust and safety assessment from the Atlantic Council's Democracy + Tech Initiative at the Digital Forensic Research Lab, the report's authors draw attention to the "clear governance challenges" facing instance administrators on open, federated systems, which mean that "each instance operator has to reinvent many of the policies and procedures of moderation for themselves"—all of which sharply increase a range of risks across

these distributed systems. As of last fall, we came to believe that a structured ethnographic inquiry into the current state of governance and administration models and structures, followed by analysis and pragmatic reporting-out of the results, is the best next step to take to reduce these risks, so that's what we've attempted to do in this project.

What we mean by the Fediverse

The Fediverse as a concept has been around in one form or another since about 2008 with the creation of an open source microblogging service called StatusNet by Evan Prodromou. Over the next few years, a constellation of social media projects coalesced around StatusNet into an interoperable network using "a bouquet of existing protocols" known as OStatus (Strype 2018). Depending on who you ask, other software that interoperated with StatusNet servers through protocols like diaspora* were also part of the Fediverse. By the mid-2010s, the shortcomings of OStatus led Prodromou, Christine Lemmer-Webber, and others to create a sort of successor protocol called ActivityPub. Unlike OStatus, this protocol was created via formal W3C governance mechanisms. Mastodon's adoption of ActivityPub in 2017 occurred at a time when Mastodon was seeing its first major increases in usage. The prospect of access to Mastodon's user base combined with ActivityPub's advantages and official W3C status incentivized other projects to move to ActivityPub as well. These days the Fediverse is understood as "a decentralized, open source, largely nonprofit ecology of bounded, linkable social media sites, apps, and services (e.g. Mastodon, Pixelfed, Lemmy), all built on the ActivityPub social web protocol" (Struett 2023). However, the "open source" and "largely nonprofit" portions of the definition have been complicated by the 2024 entrance of Meta's Threads microblogging service into the Fediverse.

The Fediverse is ever-changing in scope, and we think its current incarnation is best described as a decentralized—or non-centralized, as discussed in Section 3 below—interoperable network of social media sites, apps, and services built on the ActivityPub protocol.

What we mean by governance

The word "governance" has its roots in the same Ancient Greek term for piloting or steering that gives us "cybernetics." With this sense in mind, we take a broad view of governance, inclusive of all the socio-technical norm-setting, policy making, listening, structuring, management, and other forms of steering that are intended to keep Fediverse servers on course and afloat.

The server admins and moderators we spoke with described dozens of social and technical ways Fediverse servers are governed, some of which are extensively documented and obvious to

members and many of which are much less so. In our findings, we've tried to capture as many of them as we can, in as much detail as we can given our project's timeline.

In this report, we focus on three main areas of governance and a fourth topic that cuts across all three:

- **Moderation, or the governance of server members and content.** The social/cultural aspects of moderation are heavily entangled with the technical tools used to communicate and act on policies, so although we deal most directly with moderation in Section Two: Moderation, our detailed discussion of moderation tools in Section Five: Tooling is also relevant.
- **Server leadership, or the governance of the server and the people running it.** This aspect covers our understanding of decision-making, formal and informal team structure, how authority and responsibility flow, and how resources are chosen, allocated, managed, and sustained and is dealt with most in Section Three: Server Leadership.
- **Federated diplomacy, or the governance of relationships between a given server and other Fediverse servers and accounts.** This aspect includes federation with—or defederation from—other servers and their members and systems of cross-server information-sharing, and is covered in Section Four: Federated Diplomacy.
- **Tooling for governance, including software and financial and legal mechanisms.** Section Five: Tooling discusses tools relevant to each of the three other sections: moderation technology, use of technology to coordinate internal governance, and gaps in tooling for inter-server communication and relations.

We haven't attempted to outline a single path from "bad" or "insufficient" or "simplistic" governance to good or sufficient or sophisticated governance—nor will we suggest that more sophisticated or complex forms of governance are better than simpler ones.

For our purposes, effective governance on the Fediverse is governance that is appropriate and positive for a given server's members while remaining either positive or not actively negative for the network's broader membership. This is still inescapably subjective, both for us and for our interviewees. But room for subjectivity is also one of the Fediverse's gifts.

Because we've chosen to focus on socio-technical governance—and especially on the places where interpersonal and technical work overlap the most—we elected to exclude the most purely technical forms of server administration, including hosting, provisioning, and day-to-day technical administration of Fediverse servers. Nevertheless, we think this would be a fruitful subject for exploration—and would probably benefit from a broader and less ethnographic approach than ours.

A note on our evolving conceptions of governance: One of our research participants from Social.coop, Nathan Schneider, is also a scholar of participatory and democratic governance of online communities. Although the research summarized in this report is largely limited to the

expression and analysis of our participants' experiences, challenges, and aspirations, we're indebted to the core arguments of Schneider's latest book, *Governable Spaces: Democratic Design for Online Life* (University of California Press, 2024), which was published during our interviews and which influenced the way we understand server teams' descriptions of their practices, aspirations, and relationships to their technical systems. We recommend the book to readers interested in the governance of online communities and systems in the Fediverse and beyond.

Methods

We spoke with 16 operators of 11 teams running medium-sized servers, along with two advisors to the nonprofit organization Independent Federated Trust and Safety (IFTAS), one of whom is also a server administrator. In semi-structured interviews, we discussed the things administrators and moderators do to govern their servers, the artifacts they make, the tools they use, and the aspects of the Fediverse that they're most worried about and most excited about or encouraged by. (We've included some of the responses to the latter category as a special all-optimism section at the end of this report.)

Notably, we didn't ask directly about more philosophical points, but most of our participants situated their answers within an explicit or implicit sense of the ethics and responsibilities of server operation, and we've included those nuanced framings throughout the report.

Why we focused on medium-sized Mastodon and Hometown servers

"Medium" is subjective, but we see governance needs as something that changes as communities change size. Through our governance lens, "small" refers to any community of a size where governance needs are minimal. We see a shift as servers approach about 75-100 community members—above this size, the need for governance becomes more acute. We also specifically wanted to construct a sample that featured mostly servers with multiple moderators and with relatively thorough documentation, which kept us focused on servers with more than a few dozen members. We set our floor for server selection at 80 members, minimum.

For the purposes of our research, we consider "large" servers to be those with more than 10,000-15,000 community members; the largest server in our sample hosts just under 11,000. While studying large servers would certainly prove useful and interesting, our hypothesis was that some of these larger servers operate more like typical corporate social media platforms, albeit small ones compared to Facebook et al., and that their socio-technical interactions will probably be closer to what is already understood in depth by existing literature studying corporate social media. Other large Fediverse servers may operate more like the smaller ones we studied, but we suspect that their scale alone would nudge them toward structures and processes that might be

less useful to the operators of medium-sized servers whose needs we're attempting to highlight (and meet, in part) in this project.

In terms of monthly active users, the majority of the Fediverse's nearly 30,000 servers (29,132 according to FediDB in June of 2024) are very small. According to FediDB, only one fully federated server—Mastodon.social, run by Mastodon gGmbH—has more than 25,000 monthly active users, and that server has approximately 230,000 active users as of June, 2024. Another eleven servers listed on FediDB report more than 10,000 monthly active users; most of the remaining 29,000+ servers are much, much smaller, with many hosting only a single account. Our area of interest—servers hosting roughly 100-10,000 monthly active users—therefore targets a group of servers on the "large" end of the whole Fediverse spectrum, so our initial research question referred to "medium-to-large" servers. However, we realized early in the project that using "medium-to-large" in our description led some readers to believe that we weren't focusing on any servers with fewer than 1,000 monthly active users, so we've switched to "medium-sized" as the simplest and clearest descriptor.

Given our timeframe for this study and available resources, we further limited our scope to only servers using Mastodon and Hometown social media software. We chose these two pieces of software because, taken together, they comprise the largest bloc of Fediverse activity—and because they're where we already have pre-existing expertise. Both Erin and Darius have used Mastodon for years, and Darius maintains the Hometown software (which is a modification of Mastodon). Given our short research timeline, we chose to limit the number of systems to allow us to go deeper on server teams' experiences. There is clearly much additional work that needs to be done looking at non-Mastodon-based communities on the Fediverse, including deep dives on communities using individual software projects, comparative study across software projects, and more.

How we chose our interviewees

We approached the selection of our interviewees quite deliberately because we think that in-depth interviewing produces valuable insights that don't come up in more shallow engagements, but also requires the construction of a very short list of participants, and we wanted to get that list right, for multiple values of "right."

Because our aim in this report is to help better distribute the expertise—and multiplicity of approaches to active governance—found in the teams running Fediverse microblogging servers, we limited our scope to focus on servers that have at least two team members involved in administration and moderation, to servers that attract members from outside immediate friends-of-friends circles, and to servers that take active responsibility for governance.

In the plainest terms, that means we didn't make attempts to interview people running servers with fewer than 80 users, servers with only one admin/mod, and servers that are uninterested in (or hostile to) governance and moderation as aspects of online community. We don't doubt that there

are various insights to be gained from conversations with people in those groups, but it wasn't what we were after.

Working from a long list we brainstormed together and supplemented with a surprisingly productive informal self-nomination process on our personal Fediverse accounts, we developed a series of weighted selection criteria to ensure variety across:

- number of active users
- location of server/admin team
- server focus (general, regional, topical, focus on specific minoritized communities)
- language
- governance structure (BDFL, co-op, etc.)
- legal/entity structure (informal, LLC, nonprofit, etc.)

Once we'd built a balanced shortlist using those criteria, we contacted the operators of each server via email and Mastodon direct message to ask them to fill out a lightweight, privacy-protecting interest survey, and once we'd received interest forms from more than six server teams, we invited participants—some for full-length interviews, and some for lighter conversations.

Although we're unaffiliated with a research institution and therefore have no institutional review board, we wrote a privacy and consent document for all participants and sent this document to each operator before our conversations. We also began each interview with a discussion of potential risks of participation and the range of possible redactions and obfuscations we are able to provide to ensure that each participant was comfortable with the level of identifiable detail present in our findings.

Also, as we noted at the beginning of every interview, our aim in conducting this research is to represent the experiences and perspectives of our participants accurately and authentically. We provided all interview excerpts for publication to our participants for review and redaction, and have redacted identity in certain places throughout our reports to allow our participants to speak freely. (Quotations have also been lightly edited to remove many of the verbal fillers and false starts always present in oral interviews.)

This overall approach resulted in a sample that is—as all samples are—imperfect, but still more representative of the broader Fediverse experience than we'd have achieved otherwise.

Our final set of 11 server teams, each of which contributed between one and three participants, includes:

- eight servers with between 80 and 2,000 monthly active users and three servers with between 2,000 and 11,000 monthly active users at time of interview;

- a range of governance (BDFL, cooperative, informally participatory), legal (non-profit, LLC, project of institution, no entity), and registration (open, closed, open by application) structures;
- six region-focused servers (two in Western Europe, three in North America, one in South America), three of which are primarily non-Anglophone; and
- two topical/subculture-focused servers, two servers aligned with queer and/or trans communities, and one academic-affiliated server.

We also spoke with two people affiliated with IFTAS (Independent Federated Trust and Safety), a non-profit organization conducting research along similar lines to ours, among other projects intended to make the Fediverse a safer and more trustworthy place.

The approach we've taken in building our selection and interview processes—which we'd characterize as deliberate, transparent, sensitive to individual needs, and reciprocal—also necessarily shaped our findings.

What we missed

We regret that despite reaching out to several additional servers outside of Western Europe and the US and Canada, we couldn't get more of those servers into our sample—several servers we approached had MAUs below our minimum cutoff, one was in the process of shutting down, and several others didn't respond to our outreach. The same was true for servers focused on specific ethnic and racialized communities: the servers we found were out of our range, and those near enough to potentially make an exception for didn't respond to our inquiries. **We think future research focused on small servers would be more successful in capturing insights from these locations and communities**—our focus on a variety of governance structures and on servers with multi-person admin teams and more than 80-100 MAUs was extremely helpful in bringing the governance information we were seeking to light, but necessarily exerted shaping effects on our sample.

Due to tightly constrained translation resources, lack of relevant language skills on our two-person team, and our abbreviated timeline for the project, we didn't approach any teams in the large cluster of Japanese-language servers. The contexts around this cluster are also highly specific and require more attentive treatment than we could offer in our study of governance models. We think this is a rich area for future research.

We're grateful to the many participants for whom English is not a first language and who spoke and corresponded with us in mostly English anyway!

How we assembled this report

Erin ran point on detailed research questions while Darius set up our systems and built out tooling. After collaboratively running the interview process with our participants and reviewing our findings, we divided the report into sections: Erin led analysis on major themes, overall risks and recommendations, moderation, the cultural side of server leadership, and federated diplomacy; Darius led analysis on software/tooling, legal questions, and financial concerns.

This primary report is accompanied by two additional documents intended for specific readerships: **Fediverse Governance Opportunities for Funders and Developers** (for funders and developers) and a **Quick-Start Guide to Fediverse Governance Decisions**, for people interested in founding, running, or joining Fediverse server teams.

Brief glossary

This glossary is intended to define terminology as you will see it used in this paper. These definitions are not meant to be global or normative—they're just references for internal consistency and convenience. For our purposes...

- The **Fediverse** is a decentralized interoperable network of social media sites, apps, and services built on the ActivityPub protocol.
- Fediverse **servers** are websites that connect to other websites using ActivityPub. They work like this: A person points a web browser to "social.example.com" and sees a welcome page that says something like "Welcome to Example Social! Click here to create an account." That person signs up for an account. They now have an account on the server, from which they can follow accounts on many other servers in the Fediverse. (Server and **instance** mean the same thing and are used interchangeably by many of our study participants.)
- **Members** or **users** are the everyday people who have accounts on a given server, but aren't server operators. (We tend to use "members" when writing about the human aspects of governance and "users" when writing specifically about software.)
- **BDFL** is an acronym that stands for "benevolent dictator for life". This is a term widely used in open source communities and is a tongue-in-cheek reference to what is likely the most common governance model in open source software: a single individual who is tied to a software project and gets final say on all decisions regarding the software. **BDFN** is a more recent coinage: the "benevolent dictator for now," to denote administrators who have expressed the willingness to to step down, hand off the project, or move toward a more participatory model at some point.
- A **server team** is a group of people who are responsible for running a server. Most of our participants from server teams were administrators, moderators, or both, although some were advisors, board members, or members of cooperative working groups.

    - An **administrator** is a person with privileged access to information and control over the configuration of a server. They hold the most material power on a server. While an administrator may be beholden to membership votes and so on, they do hold the metaphorical keys and are entrusted by all the members of a server to behave responsibly.
    - A **moderator** is a person whose duties include but are not limited to filtering content, setting norms, enforcing a code of conduct, and adjudicating interpersonal problems on a server. A moderator has more material power on the server than a typical member (for example, the ability to delete anyone's posts), but not as much material power as an administrator (who could, for example, delete the entire server). Many administrators are also moderators, but the two positions do not necessarily overlap.
    - We also use "**operator"** as a generic term for someone holding any position on a server team.

- **Federation** is the act of connecting one Fediverse server to another. It opens up a sort of content firehose between the two servers and each becomes aware of publicly available activity published on the other. For a user on one server to talk to a user on another server, the two servers must be federated. **Defederation** is the opposite: a severing of this tie so that content no longer flows between the two servers.
- **Federated diplomacy** is the governance of relationships between a given Fediverse server and other Fediverse servers and accounts. This aspect includes federation with—or defederation from—other servers and their members, and systems of cross-server information-sharing.
- **Limit** is a Mastodon-specific term referring to a type of moderation where an account is hidden from any users that don't currently follow it. No connections are severed but there is no discovery of that account from users on the server that has limited the account. This is roughly equivalent to a "mute" on other social media sites.
- **Suspend** is a Mastodon-specific term referring to a moderation action where an account is effectively deleted, data purged, and any messages to or from that account are rejected. This is roughly equivalent to a block or ban on other social media sites.
- **Adversarial behaviors** and **adversarial servers** are those whose actions run counter to the underlying principles of thoughtfully governed Fediverse servers. "Adversarial" is necessarily a subjective term, and here includes spam, scams, coordinated harassment, covert influence campaigns, and other behaviors that constitute abuse of the network's social and technical affordances.

Special thanks

For the invaluable assistance as research participants, reviewers, advisors, translators, or more than one of those roles, we thank Johanna B., Larissa Babak, Tim Bray, Bumblefudge, Renato "Lond" Cerqueira, EverydayMoggie, Kathleen Fitzgerald, Eduardo "Flancian" Ivanec, Robert Gehl, Phil Siino Haack, Ashkan Kazemi, Jaz-Michel King, Kyle Kingsbury, Samantha Lai, l4p1n, Manon Marchand, Katharina Meyer, Jon Pincus, Pine, Evan Prodromou, Quintessence, Ryan Randall, Nathan Schneider, Moritz Steiner, the participants who preferred not to be named in print, and the folks at the Digital Infrastructure Insight Fund for making this work possible.

Section One: Overall Observations

Our aim in this report is to document **how governance happens** within our sample of Fediverse microblogging servers and to identify several modes and methods of governance that work well for our participants, to discuss common threats to effective governance, and to make brief recommendations that either emerge from the practices of the servers we studied or which our interviewees mentioned as possible ways of handling risks.

The bulk of this report (Sections Two through Five) focuses on detailing the governance practices, tools, and challenges our participants described. This section, in contrast, lays out our **broad observations about the character of governance practices** we encountered, the **governance-related risks** our participants discussed, and the **potential mitigations** that emerged from our conversations.

1. The big themes

In governance terms, the Fediverse is best conceptualized not as a social platform or network, but as a social component of the open web, with all the benefits and drawbacks this entails.

Media reports and scholarly approaches often position the Fediverse as an ungovernable version of the centralized social platforms that have become such powerful agents in online and offline life, but we think a simpler and more accurate framing is that **the Fediverse operates according to the pre-platform logic of the open web**.

On the Fediverse, server operators can choose which entities to maintain federation relationships with and which entities to exclude, but the network has never had an authority capable of accepting responsibility for any given server's behavior or existence—much in the way that one website (Wikipedia, perhaps) is not responsible for the behavior or existence of another website (4Chan, let's say). Responsibility for the removal of illegal content on the Fediverse falls first on local operators, but ultimately rests with local law enforcement—as it does for other websites on the open web.

It may be useful to think of the Fediverse as not truly *de*centralized, but, in digital governance scholar Robert Gehl's formulation, *non-centralized*, having never been centralized to begin with and therefore not being subject to a purely retrograde *re*centralization. **From this angle, the Fediverse is more similar to a series of pre-platform web forums that can choose to talk to each other than it is to any of the centralized platforms that defined the past 15 years of social media.**

To take the antisocial drawbacks of this kind of system first, this means that potential Fediverse members seeking a place to practice speech that is widely prohibited under local law, such as child

sexual abuse material (CSAM), nonconsensual intimate imagery (sometimes referred to as "revenge porn"), extremist/terrorist recruitment material, or material classified as hate speech in various jurisdictions, will often be able to find homes on Fediverse servers that accommodate them. The same is true for people and groups who want to enact the kinds of harms that aren't illegal in most jurisdictions, but which are prohibited by many centralized social platforms, like network abuse, covert coordinated influence campaigns, and speech that denigrates others based on protected categories often including race, ethnicity, gender, gender identity/trans status, disability, religion, age, and place of origin.

Servers that host these kinds of content—whether willingly or through neglect—illustrate the crucial difference between the governance of centralized "walled garden" platforms and governance on the Fediverse. Where a centralized platform can seek to identify and suspend accounts and networks of accounts posting illegal, abusive, or otherwise impermissible content according to their terms of service, Fediverse servers have no such power. Instead, Fediverse servers act locally, and frequently coalitionally: For example, most mainstream Fediverse servers defederate from those hosting unquestionably illegal content, and as a result, "worst of the worst" Fediverse servers are in practical terms walled off from most mainstream Fediverse servers. This doesn't *delete* the servers hosting illegal/abusive/extremist material—like the many other sites hosting illegal and extremist material on the open web, bad Fediverse servers remain online unless or until taken down via local law enforcement or through appeal to their technical hosting providers.

Importantly, the defederation of servers, individual accounts, and individual messages on the Fediverse isn't limited to illegal content—or even to the kinds of content often prohibited by centralized social media platforms. Because of Fediverse servers' ability to defederate from other servers at will, people seeking refuge in which to communicate freely about topics that commonly make them targets of coordinated harassment on centralized platforms can, in theory, find homes on Fediverse servers that sensitively accommodate their communication needs while aggressively defederating from servers willing to host the people devoted to attacking them. In this way, **Fediverse governance *can* be much more focused on local norms and community needs than any large-scale centralized platform.**

Medium-sized servers within a non-centralized, federated system offer uniquely supportive conditions for community self-governance according to local norms.

In the simplest terms, Fediverse servers need not attempt to be all things to all people, and can instead focus on becoming **the right thing for a given group of people**. The social and political facts of the ActivityPub-based network's structure—in particular, the ability of members to choose their experience from among many Fediverse experiences, and the ability of server teams to defederate from servers that behave in ways they consider destructive to their own members—make the ecosystem suitable for the construction of many kinds of community

experiences, including those centered on frequently censored or targeted communities, as well as experiments in participatory and democratic exercises of power.

Taking advantage of the unique opportunity offered by the Fediverse requires two things, at minimum:

First, it requires **the creation and sustenance of many smaller and medium-sized servers capable of putting forward and enforcing coherent statements of their values, policies, and commitments to their target communities and of governing their servers according to their communities' needs and norms.** The teams we spoke with represent servers that have achieved—or are well on their way toward—this level of service provision for at least some communities and members. Many are also exploring more formalized methods of governing themselves as servers and teams, both as a means of achieving greater organizational (and therefore also technical) sustainability and in service of the ideals of self-governance.

Second, it requires **the development of better ways for new, potential, or dissatisfied Fediverse members to identify servers that meet their governance needs**—which will first require helping them to understand what their needs actually are and what factors to consider as they evaluate server governance across the Fediverse—and move to those servers with maximal ease and minimal loss, no matter their level of technical sophistication.

We think it's important to acknowledge that in practice, some communities have flourished more than others on the Fediverse, and this has in turn shaped the Fediverse's current userbase. In a particularly stark example, as we prepare to publish these findings, a new surge of discussion is taking place on and around the Fediverse about the many negative experiences Black members (and moderators, developers, and admins) have had and continue to have on the network—a Fediverse conversation that occurs with regularity but without resolution.

Our research suggests that there are significant gaps to be filled in the tooling and resources available to server teams using Mastodon and Hometown, in particular—as highlighted in "High-level recommendations" below—but that the structural promise of the Fediverse is real, and that the benefits it confers can be made available to many more people in many more places *if* these socio-technical gaps can be filled.

Medium-sized Fediverse servers can offer high-touch, context-sensitive, moderation that differs sharply from that of central platforms.

The server teams we spoke with have varying moderation ratios, but many provide more than one moderator per 1,000 members; **the most lightly staffed server provides one moderator per approximately 1,800 members, and several provide at least one moderator per 100 members**.

To put the above figures into context, in 2020—the most recent year for which we were able to find statistics—Meta employed or (mostly) subcontracted about 15,000 moderators to moderate content across both Facebook and Instagram, according to a report by the NYU Stern Center for Business and Human Rights. That same year, Facebook had 2.8 billion monthly active users (MAUs). Meta doesn't publish official Instagram user numbers, but according to CNBC reporting, Instagram had approximately 1 billion MAUs in 2018 and 2 billion MAUs in 2021. Even if we use the older, lower number, Meta would have been providing only **one moderator per approximately 250,000 active users** across its two largest platforms in 2020.

For the thoughtfully governed, medium-scale servers represented in our sample, it's possible to maintain a dramatically better ratio of moderators to active users even with a handful of moderators. Equally importantly, it's possible to build moderation teams that are representative of a given community, and which are focused on moderating according to the specific concerns and norms of that community, rather than on enforcing one-size-fits all "community guidelines" delivered by a centralized organization.

Very large Fediverse servers may be able to provide similarly attuned moderation by aggressively scaling up their moderation teams, but these servers were outside the scope of our research. Anecdotally, several server operators we spoke with noted that most of their members' day-to-day reports about spam or harassment are about bad behavior by members of very large, lightly moderated servers.

Sustainable governance results from making the right set of interconnected socio-technical choices for a given server.

In our interviews, we heard teams describe ways of operating servers ranging from fewer than 80 to more than 10,000 active members, with several different registration models (open, closed, and variations on invite-only) with moderation teams of various sizes, with documentation and rulesets ranging from short and simple to voluminous and complex, with a range of legal and financial structures, and with varied approaches to deny-lists and other software-based tools.

We had expected to be able to sort the servers we worked with into a few discrete groups and then discuss the servers as exemplars of the various groups. Instead, as we synthesized the results, it became clear that our sample set—which, again, included only teams with the excess capacity to send members to speak with us!— represented **many different ways of approaching the same trade-offs**.

It is *possible* to divide the servers we studied into groups along various axes, and in the detailed findings below, we do so when it makes sense for a given sub-topic, but at the high level, we found it more illuminating to consider the way specific governance choices increase or reduce pressures, which can then be further up- or down-regulated by other governance choices.

Examples from our findings:

- Open registration tends to result in larger active member numbers and more unpredictable sign-up numbers, even for nominally regional or topical servers, both of which increase risk surface and moderation volume. Operators of open-registration servers can compensate for these factors by scaling up moderation teams, publishing more detailed documentation, outlawing or restricting more account types (commercial, institutional, etc.), and assigning resources to actively investigating and defederating from ungoverned or adversarially governed servers that produce a high volume of problematic content or behavior.
- In contrast, moderated or closed registration allows server operators to keep sign-up numbers lower and/or filter out applicants who seem unlikely to contribute to the community culture a server seeks to provide, both of which allow operators to provide active, high-context moderation with smaller moderation teams and simpler processes and docs.
- Medium-sized servers with a simple leadership structure can often cover operating costs via donations, while server teams invested in more participatory or democratic forms of server leadership or in more formal legal structures may choose to require dues from members and/or to seek private funding (from server operators or others) to cover the cost of establishing legal entities—and in the case of more democratically governed servers, to strengthen the relationship between a server and its members.
- Community context should guide server policy at every level. Servers that seek to provide a home for members of marginalized communities or people who represent one position within a politically contentious landscape generally maintained smaller server sizes and more restricted registration, and discussed their approach to moderation and federated diplomacy as more aggressive, contextual, and high-touch, while more general-interest servers tended to describe their moderation responsibilities in terms of providing the most freedom while reducing or eliminating obvious harms.

The *diplomacy* aspect of Fediverse governance is critically important to the successful operation of servers, but remains largely opaque.

The relationship between any two Fediverse servers is, essentially, a diplomatic one between two sovereign powers. The team running one server, no matter how large or influential, can't force any other server to take a given action. The threat of defederation ("limiting" or "suspension" in Mastodon's terms) by one server or an informal coalition of servers is the only built-in lever in the Fediverse for the cross-instance exercise of power.

Given that full defederation cuts connections between the two servers' members and prevents those members from re-establishing them unless they switch servers or the two servers re-federate, this diplomatic layer of governance has a significant effect on server members' experiences: People on servers with a reputation for hosting spammers and trolls, for example, will find themselves cordoned off from many Fediverse servers, while people on more mainstream but lightly moderated servers may be unable to connect with members of servers that aggressively defederate from servers that don't moderate according to their more restrictive norms.

Especially for novice Fediverse users, these dynamics can be confusing or even invisible, particularly given the near-total lack of public communication outlining most servers' defederation policies. This is an especially challenging factor for would-be server members trying to sort out which Fediverse server to choose, since defederation has strong effects on the way a server's members will experience the Fediverse, including how much abuse, harassment, hateful or violence-inciting speech, and spam they're likely to see.

The situation is made more complicated by the lack of in-system communication channels between server teams using Mastodon and Hometown. Although they have strong controls for moderating their own members and managing individual messages posted by their own (or remote) members, the only way for Fediverse server operators to interact with other server teams is by communicating with them informally using side channels or Mastodon/Hometown direct messages—or by limiting or suspending federation with the team's server.

As a result of all these factors, server administrators and moderators doing active governance make decisions every week (or every day, for higher-volume servers) about whether and when to limit or suspend federation with other servers and individual members of other servers, but their decision-making processes and the policies behind them are often unclear.

Fediverse server teams are reliant on the ecosystem's relative obscurity and small size to handle adversarial behaviors and campaigns; what works now probably won't work forever.

As noted above, most server operators we spoke with keep moderation workloads under control by reducing attack surfaces using technical tools including moderated or closed registration and the maintenance of thorough, up-to-date defederation lists. These choices work together to free up moderator time for genuinely complex situations, which is especially important for small volunteer moderation teams.

That said, many admins and moderators expressed a sense that their approaches and processes are largely working well for now, often with a sense of anxiety about the way potentially rapid growth in the Fediverse could result in the loss of the security-through-obscurity benefits the network has retained to date.

The kinds of threats Yoel Roth and Samantha Lai refer to as "collective security risks" in their recent paper, "Securing Federated Platforms: Collective Risks and Responses"—sophisticated spam attacks and coordinated covert information campaigns, in particular—do pose a looming threat in many admins' and moderators' minds, though it's perhaps noteworthy that these elements have not yet become major aspects of the experience of most teams we spoke with.

It's also possible that, as Roth and Lai note, the Fediverse's lack of algorithmic acceleration mechanisms and built-in financial incentives will continue to exercise a protective effect at the structural level even as the network expands. We think this will almost certainly be true to some extent, but that it's difficult to predict the dynamics of a much larger and therefore more target-rich Fediverse.

If the Fediverse continues to grow, server operators will require more sophisticated ways of identifying and rooting out unwanted content and campaigns to maintain a healthy environment for both their members and their moderators. Given the Fediverse's history of fierce independence and mistrust of surveillance—and the ambivalence expressed even in our governance-friendly sample toward tools like widely shared blocklists— we think platform-style centralized telemetry is unlikely to be an acceptable solution for the majority of server operators.

Fediverse governance as we encountered it in our research conversations is emergent, unevenly distributed, and often reactive.

Although it often shares concerns with centralized platform "trust and safety" practices and systems, the governance of Fediverse microblogging servers is fundamentally unlike those practices and systems. Fediverse governance emerges from the combination of the unique affordances and limitations of the ActivityPub protocol and the software built on top of it, the many experiments in internet community management at multiple scales that we've seen to date, and the various evolving consensus understandings of the roles governance should play within the diverse communities that have participated in the Fediverse.

From the perspective of centralized platforms, the **emergent nature** of Fediverse governance can seem chaotic and even irresponsible or dangerous. We consider it a central characteristic of the Fediverse that will continue to define the shape of current and potential risks—and of Fediverse-appropriate mitigations—in both negative and positive ways. We therefore expect effective solutions to commonly experienced problems to emerge from many corners and collaborations, and to be multivalent rather than monolithic.

Thoughtful governance is far from being an innate quality of the Fediverse. Many—probably the majority of—Fediverse servers with more than a few members are largely unconcerned with governance in the way we discuss it here. Our sample servers include several of the most

governance-conscious teams on the network and aren't representative of the Fediverse as a whole, or even of medium-sized Fediverse microblogging servers. But multiple thoughtful modes of governance have nevertheless emerged on the Fediverse, and we've been fortunate enough to be able to speak with many of the people practicing them.

Even within our sample, **the nature of governance structures and modes varied widely**: nearly all teams we spoke with had sturdy and carefully thought out moderation processes, norms, and teams in place, but far fewer had server leadership structures that extended beyond informal Benevolent Dictator for Life (BDFL)/Benevolent Dictator for Now (BDFN) open-source defaults, and even fewer had clear policies or practices for evaluating the kinds of communication with and decisions about other servers that we categorize as "federated diplomacy." And most of even these exceptionally well-prepared teams told us that their approach to at least some aspects of governance is still evolving as their server (and the Fediverse) matures and expands.

A few server teams we spoke with launched with extensive documentation and policy, but most developed their policies and processes **on the fly and as needed**, often based on a simply stated set of shared values. Even for the teams who began with more process and documentation, specific incidents like the early 2024 spam wave, the development of the Bridgy Bluesky bridge, and especially the federation of Meta's Threads product, have pressed teams to more publicly frame their underlying philosophies and commitments to their users, often with greater consultation with their members than was their previous norm.

Note: In this report, we draw on the usage of "emergence" that adrienne maree brown references and reweaves in *Emergent Strategy: Shaping Change, Changing Worlds* (AK Press, 2017) and which runs back through the long history of complexity science and its predecessors lucidly summarized in Peter A. Corning's "The Re-emergence of 'Emergence': A Venerable Concept in Search of a Theory," (*Complexity*, Vol 7, No. 6, 2002). The emergent character of governance on the Fediverse deserves a lengthier discussion, but for the purposes of this report, emergent systems are those which are more than the obvious sum of their parts, which develop complex forms out of simple conditions and constraints, and which result from interaction between heterogeneous actors and systems.

2. The risks

In the fall of 2023, we proposed the line of research discussed in this report because of our own individual senses of current and impending risks to the Fediverse and its members, and these senses necessarily shaped our conversations with server admins and moderators. That said, our conversations with server teams led us to a somewhat different understanding of which risks are in the foreground for the people running thoughtfully governed Fediverse microblogging servers, and which are present as variously intense forms of anxiety-producing background radiation.

For our sample—the various characteristics of which we outline in the "Methods" section above—the risks we heard about fell into several broad categories.

Class 1: Risks framed by our participants as manageable

Class 1 risks include those that most of the server teams we spoke with consider to be solved well enough—or solved "for now"—using the practices and processes they related to us.

Some of the teams we spoke with are still struggling with Class 1 risks, and the vast majority of small and medium-sized (and some large) Fediverse servers are likely to struggle with most or all of them at some point in the life of their servers. It's our hope that this report and its lighter-weight companion documents will help distribute the expertise server teams have shared with us so generously.

- Internal risks
    - bus factor for admins and moderators
    - basic financial stability
    - governance that fails to meet members' basic needs
    - foundational legal liability
    - lack of training for moderators
    - autocratic or brittle server leadership that can't respond to community concerns
    - vulnerability to both malign and well-intentioned but inexperienced and/or reactive moderation team members
    - moderator burn-out
    - technical instability
- External risks
    - unsophisticated spam campaigns
    - unsophisticated trolling, abuse, and other obviously adversarial behavior

Class 2: Sources of moderate unresolved frustration or anxiety

Class 2 risks represent immediate gaps that server teams flagged in our conversations. Filling these gaps will allow them to stabilize, improve, and in some cases expand their work. Most mitigations for Class 2 risks will involve un-flashy work across both cultural and technical domains.

- Moderation
    - time-consuming, heavily manual moderation tools
    - difficulty of onboarding new members
    - clunky and insufficient appeals tooling
    - inability to moderate in culturally attuned ways for broader ranges of members (examples included moderation across languages, geographic region/regional norms, race/ethnicity, gender and gender identity)
- Leadership
    - the liabilities of informal (no formal entity) server team structures
    - the limitations of top-down server governance
    - the limitations of highly consultative server governance
    - the high cost of formalizing non-profit entities
    - the complexity and increased social risk presented by non-profit boards
    - lack of exemplars for more consultative and participatory forms of governance
- Diplomacy
    - lack of ability to communicate easily or well with other server teams
    - rarity of receiving responses to reports from other server teams
    - contentious inter-server relations leading to over-blocking, bad feelings, and/or situational (rather than rule-based) decision-making
- Tooling (not already captured above)
    - lack of readily-available tools to detect and report illegal content
    - lack of options for federation—the all-or-nothing approach to federation taken by most Fediverse core software leaves much to be desired especially for smaller communities
    - lack of repositories of legal and financial guidance for server operators

Class 3: Sources of broader and more intense anxiety

Class 3 risks represent potentially existential threats to the Fediverse as it's understood by the teams we spoke with, and would benefit from collective and multifaceted approaches across social/cultural and technical domains.

Risks in the Class 3 list are largely not discussed directly in the body of our findings because our server teams aren't able to address them directly in their course of their daily work. These Fediverse-wide potential risks were often deferred to the end of our interviews with admins and moderators, both because they're less closely tied to the day-to-day work of server operation and because the work of mitigating will require coordinated efforts beyond the abilities of small teams, or even of the developers of Mastodon or Hometown. Nevertheless, they represent real and often potent anxieties for the people maintaining Fediverse servers.

- Ecosystem-scale loss of momentum/Fediverse die-off
    - difficulty in finding a server for potential Fediverse members
    - server longevity (the lack thereof)
    - overwhelm by well-funded alternatives
    - adversarial discourse becoming overwhelming
    - fragmentation of communities across non-interoperable decentralized systems
    - potential for corporate capture
    - lack of financial and human-effort sustainability across less conscientiously organized servers (including financial burden of duplicative media hosting)
- Socio-technical vulnerabilities
    - vulnerability to more sophisticated coordinated adversarial campaigns (spam/scams, covert influence operations)
    - bridges to other social media ecosystems increasing attack surfaces
    - impending acceleration of LLM-powered adversarial campaigns
    - lack of comprehensive understanding of attack surfaces among members and admins
    - the evaporation of critical infrastructure like the Open Collective Foundation
    - XZ-style vulnerabilities to attacks produced by an under-resourced software development ecosystem

In this document and in our accompanying brief handbook on governance, we'll outline the ways server teams have—mostly successfully—mitigated the Class 1 risks and their variously successful attempts to grapple with the Class 2 risks.

We think Class 2 and Class 3 risks present ideal targets for near-term research, funding, and development, and our concluding recommendations highlight potential lines of research and effort that may produce mitigations for both Class 2 and Class 3 risks.

3. High-level recommendations

Our conversations with server teams have allowed us to offer two kinds of recommendations: The first is a set of best practices that have emerged on various servers in our sample, and which address risks outlined earlier in this section. The second is a set of variously intense interventions to address currently unmet needs and unmitigated risks to successful Fediverse governance.

Best practices for server teams

These are actions server teams—especially teams considering setting up a new server, but also many existing teams—can take or consider today without waiting on feature development or institutional support. Based on the insights and experiences of the teams we interviewed, we would recommend that server teams:

- Consider server governance/leadership models early, before decision-making processes and tech stacks are locked in or harder to change. This isn't meant to be a roadblock to experimentation—teams don't need to incorporate a formal legal entity to think about governance structures, but especially if they're interested in forming a cooperative or other not-completely-top-down server, knowing that early on can help guide other decisions, including the selection of a technical stack. (Section Three: Server Leadership)
- Choose an account registration model carefully and with an understanding of the trade-offs of open, moderated, and by-invitation/closed registration—and the various things admins and moderators can do to mitigate the risks of more open models, including publishing clear rules and process documentation, staffing more moderators, and maintaining aggressive defederation lists. (Section Two: Moderation)
- Seek out moderators with strong on- and offline community management experience, low reactivity, and potentially first or second-degree IRL (or long-term online) connections to a relevant community, to reduce the risk of disruptive problems with the moderation team—and unnecessary stress on underprepared moderators. (Section Two: Moderation)
- Build a server team—which might include server admins, moderators, board members, advisory council members, and other roles—with broad representation from the community or communities the server is intended to host. (Section Two: Moderation and Section Three: Server Leadership)
- Create a generic, well-publicized, two-factor-secured user-facing moderation account like `@moderators@example.social` that the entire server team has access to, and establish rules for managing DMs from that account. (In Mastodon, this would necessitate also having shared email for moderators if two-factor authentication is enabled as recommended.)(Section Two: Moderation, Section Five: Tooling)
- Document plentifully and in ways that make it easy to understand the (desired) character of the server, the server team's sense of what its responsibilities are, and the processes and

guidelines in place for content and member moderation, inter-server governance, and governance of the team itself. (Section Two: Moderation, Section Three: Server Leadership, Section Four: Federated Diplomacy)

- Supplement Fediverse infrastructure by selecting (or building out) additional processes and systems necessary to support the degree of member participation, financial support, member communication, and intra-server-team communications teams need to run the server. (Section Three: Server Leadership)
- Consider doing onboarding and training for all new members of the server team, including discussions of past decisions, recusals, and preferred methods for handling complex or heated interpersonal problems. (Section Two: Moderation)
- Use specific cases and complex decisions as opportunities to refine (and document) the team's sense of its responsibilities and underlying goals/values. (Section Two: Moderation, Section Four: Federated Diplomacy)
- Consider working with volunteer or paid legal counsel to validate the team's understanding of its liabilities and responsibilities in the relevant jurisdiction(s). (Section Three: Server Leadership)
- Address overextension and burnout as quickly as possible, ideally before they happen, by building out more human, technical, and financial capacity than the team thinks will be needed. (Section Two: Moderation)
- Communicate transparently with members (and potential members) about big social and technical decisions and their implications, financial sustainability, and future plans. (Section Three: Server Leadership)
- Consider joining one or more server admin/moderator forums (ex: the Mastodon Discord for supporting members, IFTAS Connect) for peer support, resource sharing, and easier communication with other server teams.

Opportunities for addressing unmet needs and unmitigated risks

Solutions to the more challenging problems server teams discussed with us will require ambitious action across multiple levels of society, but we think a first step is to **clearly identify opportunities to contribute to the health and longevity of the Fediverse** and **the unique opportunities for self-governance and humane networking** it can provide. We also discuss each of these opportunities in greater depth in the accompanying document, **Fediverse Governance Opportunities for Funders and Developers**.

- **Better moderation tooling** along multiple axes: bulk report handling, support for collaborative and coalitional moderation, better communication channels for moderators and members, content filtering, and more. (Section Five: Tooling)
- **Core software support for shared deny-list management** (including features that ease the process of documenting and verifying reasons for a server's presence on a list) and for **easy and accessible allow-list federation** for servers that lack the resources to maintain sturdy deny-lists or which need to run in limited-federation mode to protect frequently targeted members and communities. (Section Five: Tooling)
- **Greater recognition of governance needs and trade-offs from core software projects like Mastodon**—potentially providing limited in-software governance mechanisms, or working with third parties to ensure governance tooling can be integrated via APIs or plugins. (Section Five: Tooling)
- **Better tooling for communicating with other server teams**, including potential opt-in and/or limited federation of moderation decisions of various kinds. (Section Five: Tooling)
- **Institutional or organized peer support** for server teams interested in building formal **cooperatives** or incorporating as **non-profit entities/associations**. (Section Three: Server Leadership)
- More comprehensive and detailed **how-to documentation and case studies** for setting up and running more **participatory models of Fediverse governance**. (Section Three: Server Leadership)
- More **comprehensive and transparent documentation** of subjectively successful **financial structures and sustainability campaigns**. (Section Three: Server Leadership)
- Institutional support in the form of **fiscal sponsorships or (non-technical) project hosting** designed specifically for or inclusive of Fediverse server teams—particularly pressing in light of the dissolution of the Open Collective Foundation. (Section Three: Server Leadership, Section Five: Tooling)
- Greater and more committed **participation in the Fediverse by stable institutions** including civic and governing bodies, cultural and media organizations, higher learning and research institutions, and technology and philanthropic organizations. (Section Three: Server Leadership)
- The development of a multiplicity of **collaborative institutions and coalitions focused on creating legally vetted and transparent data-sharing, research, and threat-analysis capacities** that respect the Fediverse's non-centralized character and allow server teams to opt in at varying levels of granularity. (Section Two: Moderation)

- Clear and welcoming communications that accurately **portray the Fediverse's benefits and trade-offs** and that **help potential members understand their needs and then find servers that will best match them**. (Section Three: Server Leadership)

Section Two: Moderation

Introduction

It's this total positive deviance situation where the best run servers—man, these are a really good social networking experience! And they are not the majority. Most people are not on these best-run servers.

—content moderation researcher

The moderation practices our interviewees described have many things in common with content moderation as it evolved in early web forums and eventually became broadly codified and professionalized across social media platforms; experts in human governance outside the Fediverse will recognize many familiar actions and principles. Many of the most interesting insights we encountered were those that highlighted disjunctures between Fediverse moderation as our server teams practice it and as it's practiced on large-scale platforms.

Although external groups studying the Fediverse often emphasize moderation deficits—mostly technical/surveillance capacities that central platforms possess and Fediverse servers lack—our research also pointed to several categories of both **structural moderation** (rules, policies, norms) and **interventions** (moderator actions) that are inaccessible to platforms that moderate millions or billions of users with tens of thousands of human moderators. This is notable especially given the comparatively immense technical sophistication of the technical systems and datasets available to trust and safety teams working inside central platform companies.

These positive capacities emerge mainly from two distinctive properties of the Fediverse microblogging landscape: The first is **federation itself**, which encourages the development of distinct local policy and behavioral norms across servers. The second is the **human scale** at which most Fediverse servers presently operate, which can allow even a relatively small moderation team to approach unclear and complex interpersonal situations with sensitivity and care.

Key observations

- **Fediverse moderation *can* offer humane, culturally attuned, context-sensitive moderation that far outperforms central platform offerings in its responsiveness to member needs and experiences.** Our interviewees described emergent moderation practices that offer an attentiveness necessarily absent from large-scale platform moderation. In our introduction below, we discuss the dynamics of this characteristic of Fediverse moderation.

- **A lot of Fediverse moderation work is relatively trivial for experienced server teams**. This includes dealing with spam, obvious rulebreaking (trolls, hate servers), and reports that aren't by or about people actually on a given server. For some kinds of servers and for

certain higher-profile or high-intensity members on other kinds of servers, moderators also receive a high volume of reports about member behaviors (like nudity or frank discussion of heated topics) that their server either explicitly or implicitly allows, and which the moderators therefore close without actioning.

These kinds of reports are the cleanest targets for tooling upgrades and shared/coalitional moderation, but it's also worth noting that except in special circumstances (like a spam wave or a sudden reduction in available moderators), this is not usually the part of moderation work that produces intense stress for the teams we interviewed. (This is one of the findings that we believe does not necessarily generalize across other small and medium-sized servers.)

- **Complicated decisions are unavoidable; consultation can help.** Most server teams raised the importance of identifying the moment when a specific situation (on the server or more broadly) requires broader discussion and potentially policymaking or substantive rules changes. The way moderators and admins handle these heavier conversations, including the degree to which they consult with their membership about them, seems to be one of the most important factors in defining a server's character.

  These decisions are also the most draining aspect of moderation for many of the teams we spoke with, and the availability of peer discussion (on or off the Fediverse) came up repeatedly as a way to find, stress-test, and validate ways forward.

- **Moderation begins at account registration.** Registration requirements have a huge effect on the moderation experience. Deciding what kind of account registration is enabled on a server is often the first technical/mechanical (rather than policy) choice a server administrator makes when configuring the server for the first time.

  Servers approach moderated registration (registration by invitation or application) in several ways, ranging from a requirement to personally email the server's lead administrator to an application process that includes cooperative membership dues.

  Servers with more open registration—and therefore larger memberships—tend to rely on more extensive documentation and to moderate more reactively (non-pejorative), relying on docs and rules to handle the kinds of socialization and norming that admins of smaller servers often do in individual conversations and through subtle intervention—but there are exceptions to this pattern.

- **Mod team size is surprisingly consistent in our sample.** Most of the servers we spoke with have three to five active moderators, with only one server with fewer moderators and a few with slightly more. Most teams consciously try to ensure time-zone coverage, and most also mentioned ongoing attempts to ensure coverage of multiple languages. Many discussed

their attempts to ensure that their moderation coverage included a range of racial/ethnic identities, gender identities, and cultural norms.

- **Vibes matter a lot.** Doing precisely calibrated, culturally sensitive norm demonstration and member socialization is difficult. Some servers take a high-touch approach and rely on a collaborative and interpersonally attuned mod team that can provide individualized guidance. Some servers in our sample, especially larger ones, handle this norming work more through more extensive documentation and through outbound communication like blog posts and moderation notes than via individual interactions; this approach is more hands-off, but still relies on having moderators who are fluent enough in the server's policies, processes, and philosophy to be able to act swiftly when something begins to go wrong.

- **Moderation team culture is crucial.** Building a moderation team with the right orientation and approach for a given server is challenging and slow, but getting it right is crucial for server stability and human sustainability.

The anatomy of Fediverse moderation

Our questions about the experience of moderating Fediverse servers were broad and open-ended, but we found that the responses tended to describe five main aspects of the work:

1. **Initial rule-making**, with special emphasis on **registration policies** and requirements
2. **Straightforward moderation tasks** like dealing with spam, closing irrelevant reports, and taking action on accounts and servers that are obviously acting against server policy
3. **Complex moderation work** including social guidance for individual members, consultative policy decisions about emerging problems on the server (or the Fediverse more broadly), and high-stress/high-legal-risk work like CSAM (child sexual abuse material) reporting
4. **Building moderation capacity**, including identifying the right people, training them, and working together effectively

We'll take these in turn, working from specific anecdotes and positions related by especially our core server teams, with additional commentary from the other server operators we spoke with.

1. Registration

One of the most striking themes from our conversations about moderation is how strongly registration policy (open, closed, by application) shapes both the amount and the tenor of the moderation work that follows on the server.

None of the six core teams we spoke with offer open registration. Two use Mastodon's native application process, one uses a waitlist process via Hometown, one uses an off-Mastodon process with a self-hosted form that generates GitHub tickets for the moderators to review, and two have closed/invite-only registrations. We did speak with five additional server teams to broaden our analysis, and three of those five teams do have open registration, which allowed us to explore the way that choice affects other aspects of server operation.

Many of the admins and moderators we spoke with—both in our core group and our briefer conversations—specifically noted that restricted registration keeps the back-stage experience of operating their servers manageable and the public-facing experience of being a member pleasant. In this way, we found that for most of the teams we spoke with, moderation begins at the point of registration. We have therefore grouped registration notes under the larger umbrella of moderation, though it also connects quite strongly to the structures of governance and even to the software choices discussed elsewhere in this report.

The Mastodon software user interface offers no communication to a new admin about the far-reaching effects their choice of registration mode can have on the shape of their community and the amount of moderation work they will have to do. This UI choice positions admins as "power users" who can be expected to think through the process/cultural ramifications themselves, but this is not universally (and perhaps not even commonly) true.

1.1 Registration by application

Our core server with a topical focus (kink/subculture) runs registration by a short in-Mastodon application process, and accepts nearly all applications. An admin we spoke to told us:

> *Our policy on the door is basically if you think you'd walk into a leather bar, like if you have any kind of interest, you're welcome. I reject almost no applications.... I'm not in the business of judging people's qualifications or kink. You're welcome if you're a complete newbie.*

This admin rarely needs to remove newly registered (or indeed any) users for bad behavior:

> *There have been very few people on the server who have created major conflicts or where the mod team has been like, "Ooh, we got to work with this person and, like, figure out how to either acculturate them or get them out of the server somehow." When it happens, it's usually obvious.*

Spotting those rare problem users requires attentive moderation and cultural fluency. In this case, the fluency is used to differentiate between the kinds of social/sexual play the server was established to host and abusive or extractive behavior *masquerading* as play—the latter of which would result in moderators asking the user to move to another server.

The regional (Swiss) server in our core group also runs registration by application, and requires prospective users to accept a substantial set of rules and provide some information about themselves and why they want to join the server.

Our core server that runs as a cooperative also requires an application to register and has a relatively rigorous application process that's handled outside of Mastodon itself. Applicants are asked to explain why they're interested in joining the server and how they'd like to participate in the life of the server, must set up a profile on the Open Collective platform before applying, and must agree to the server's detailed code of conduct.

When the registration form for this server is submitted, it creates a GitHub ticket that the on-call member of the moderation team reviews every day or two to accept or reject. A member who sits on both the community and tech working groups noted that because the server runs its application and registration process outside of Mastodon, the server shows on the official Join Mastodon site as "closed" to new members, which is misleading.

Our core server that focuses on a scholarly membership and is affiliated with an academic institution runs a waitlist-based registration process through Hometown, though their assessment process is relatively light. According to one moderator:

> *...everybody is tasked with watching new account requests as they come in, and just checking to make sure that these look like human beings that, you know, kind of recognize a little bit, at least, about our instance and our goals. But the vast majority of accounts that come through account requests, we approve, because we want to be as inclusive as possible. And if you don't look like a bot... likely, we're going to let you in.*

Two other server teams we spoke with outside our core group also run registration by application. One server running as a nonprofit cooperative requires an annual membership fee as well as residence in a specific country. A founder of that server noted the connection between moderated (and paid) registration and a manageable moderation workload:

> *One thing I want to emphasize is our moderation load has been remarkably light and I think having paid membership or perhaps even more generally approved membership where there's at least some human who looks at a membership and clicks the approve button makes a huge qualitative difference.*

1.2 Closed and invite-only registration

Masto.donte.com.br has fully closed registration, but allows existing members to invite new users via Mastodon's built-in invitation system. Earlier in that server's life, the admin had opened registration once a week or so, but later chose to close them (with invites open) to prevent the server from growing beyond the abilities of the trusted moderation team, and to prevent the server's demands—both technical and social—from becoming too consuming for the primary administrator.

> *I was like, okay, like, we're five at the time—I think we were five moderating ... at the time, when I first closed registrations, we were at the point where we have 500 users at the server. And I was like, okay, if all of them decide to go online at some point, and use Mastodon, we'll have about 100 users per mod. And that's already like quite a lot. So let's close down, wait a little bit.*

The admin expressed that closed/invite registrations are currently working well, though they didn't rule out the possibility of opening them again in the future.

A moderator at the server Wandering Shop described their experiments with invite codes:

> *We've gone back and forth on a couple things and found a little bit of an awkward thing that works well for us, which is the admin or I will generate a 100 user-invite code every week. And we put a time limit and a user limit on it and then post it in our announcements.*
>
> *So that anybody who's on our server can grab it and share it. And we have just asked on your honor, don't repost it publicly. So that we don't get like we had one incident where an author had sort of generated an open invite code and flooded us with like, just posted it on her web page or something, and said, "Come join me here" not understanding that you don't have to be on the same Mastodon server, first of all. And second, like that got us a whole flood to deal with and it was like, "Okay, no, we're not doing quite open signups right now!"*

The same moderator noted that the invitation process had the secondary effect of making the process of bringing new members into the community more participatory:

> *...we tried "register but with approval," but that also proved to get us a whole bunch of spam registrations. And it left one or two administrators having to try and identify and look up every person. With the invite code that everybody can share, we control the code, it does expire, it's limited, you can't sign up more than 100 people, and it's really brought the community into managing who joins the community in that. My joke about it was we want to be the worst kept secret handshake. You know, everybody can hand the code out to their friends or their family or somebody they met in a bookstore, we just want to have that connection to build the community. And it seems to have worked really well.*

The small US-based regional server we spoke with also has closed/invitation-only registrations, and includes a note on the server's "About" page detailing a short process for emailing the administrator to apply for an account, with an emphasis on demonstrating shared values with the server's existing user base, in addition to being based in the Minneapolis-St. Paul area.

The admin of that server spoke about their expectations for prospective members:

> *You know—it doesn't have to be a thousand word essay. But we work in text, right? Like posting is mostly text. So I need you to be a good demonstrator of that stuff.*

1.3 Open registration

The three teams we spoke to that offer open registration are also the three largest servers we engaged with, ranging from about 4,000 to about 11,000 monthly active users.

SFBA.social, which focuses on the San Francisco Bay Area, tends to permit people who live outside the target area to maintain accounts, but is considering blocking registrations from specific countries that generate disproportionate numbers of spam accounts. This server's admin team also tries to ensure that despite their relatively relaxed registration policy, the server still has a regional flavor. One aspect of this work is the careful definition of allowed account types:

> *...we spent a lot of thought on how to make it feel like a very regional instance, right? For example, when it comes to companies posting stuff, right, we are very strict. It should be companies that are tied to the Bay Area in some regard, right? Either it's a local pizza shop, that's fine. But if it's a multi-million dollar company who happens to have their seat in the Bay Area, maybe not. And also in terms of advertisements, what they're allowed to post. Yeah, they can say they have a special pizza tonight. But if they post this like every five minutes, then no, right? So we have pretty strict guidelines of what companies are allowed to do and whatnot.*

Another server with open registration maintains detailed documentation on allowed account types and specific behaviors that are permitted and prohibited for, e.g., organizational accounts, which allows them to filter their membership to a degree without requiring an application process.

An admin for Piaille.fr, a server focused on French-speaking users with about 7,000 monthly active members, noted that it's impossible for them to know all their users personally in the way that can happen on a closed-registration server:

> *I think we have more trouble regarding the fact that the instance is open to registration because we suffer from bot waves, et cetera. That can be quite tiring, I'd say. Aside from that, moderation is kind of a regular task we have to do. It's not that harsh and it does not require such an amount of work in the end.*
>
> *[…]*
>
> *I think maybe one specificity…is we have thousands of active people whereas some other instances they have maybe 500 people and they are closed registration. You maybe can know everyone inside your instance. Whereas for me, I can't….I think the most difficult thing is that everything is new and there is nothing or no one you can base yourself on. I think we are the biggest instance in France. We have to figure out ourselves what to do sometimes and how to engage people …*

The lead admin of the largest server on our list, Hachyderm, spoke extensively about the things their team does in terms of documentation, active moderation, and engagement around norms and behaviors, to manage the community experience on a server with more than 10,000 members—that admin's comments were especially relevant in our below discussions about complex decision-making and the formation of mod teams.

2. Rules and guidelines

All the server teams we spoke with maintain at least a simple set of public rules for their servers, and many maintain much more extensive documentation about their server's character, norms, and governance. A few servers also have private documentation for their moderation teams and other people involved in the server's operation.

2.1 Documentation types and links

Documentation we reviewed—both within Mastodon/Hometown "About" pages and on external sites—included:

- Server rules
- Detailed explanations of server rules
- Codes of conduct
- Allowed (and disallowed) account types
- Lists of explicitly encouraged norms
- Lists of explicitly forbidden behaviors and actions
- Lists of consequences assigned to specific breaches of rules/codes of conduct
- Descriptions of/instructions for participating in appeal processes
- Guidance on how and when to report social problems
- Conflict resolution guidance
- Moderation beliefs and commitments
- Blog posts describing servers stances, policy changes, and ongoing discussions
- Forum threads discussing (and in some cases voting on) policies and proposed changes
- Guidance on how to use Mastodon/Hometown features (aimed at newer users)

Mastodon/Hometown "About" pages from our server teams:

- CoSocial.ca
- hcommons.social
- Hachyderm
- Masto.donte.com.br
- mspsocial.net
- Piaille.fr
- SFBA.social
- Social.coop
- tooting.ch
- Wandering Shop
- woof.group

Selected additional moderation-related documentation from the server teams we spoke with:

- Hachyderm Moderation Actions and Appeals Process
- Hachyderm Blocklists information
- Hachyderm Moderator Covenant
- Hachyderm guidance on making reports and interacting with moderators
- Hachyderm process for requesting exceptions and rule changes
- Hachyderm Rules Explainer
- Hachyderm Sexual Content policies
- Hachyderm Monetary Posts policies
- Hachyderm Account Types guidance
- Hachyderm Mastodon Welcome/User guidance (includes notes on accessibility, content warnings, hashtags, and mental health)
- hcommons.social documentation (includes Server Rules, Code of Conduct, Encouraged Uses, Bannable Behaviors, Moderation Policy)
- SFBA.social Code of Conduct
- Social.coop Member Code of Conduct v3.1
- Social.coop Welcome/Join page
- Social.coop Reporting Guide
- Social.coop Conflict Resolution Guide
- Woof.group documentation (includes Code of Conduct and New Users Guide, and other docs)

2.2 Rule-making as moderation

To state the obvious, all the server teams we interviewed have posted rules, and even the shortest rulesets cover foundational principles for civil discourse, many in the form of prohibitions against harassment, incitement of violence, and identity-based discrimination or abuse. According to one server admin

> *We want to be as general and inclusive as we can. On that matter, we go as far as the Karl's Popper paradox of tolerance allows us, that means we won't allow racism and oppressive behaviors.*

Some servers emphasize positive norms, rather than prohibitions, most notably our topical server focused on subculture/kink, whose rules include this caution: "Be kind![...] We're trans-inclusive, body-positive, and anti-racist here." (These more positive statements are backed by more extensive conduct prohibitions elsewhere in the server's documentation.)

Some admins we spoke with drew an explicit connection between the specificity of their documentation and their desire to ease collaborative moderation. The founder of a regional server focused on Brazilian members related the process of trying to make group moderation easier by

codifying more decisions as written rules outlining specific consequences for specific breaches, allowing individual moderators to act quickly:

> *We tried very early on to come up with some more objective rules, in a way, so that…anyone could take either a moderation issue or even something they see on the timeline, and act on it.... You don't have to discuss for every specific situation. So our rules are a bit more like, "If you do this, there's that, and if you do that, there's that."*

Another admin spoke about writing relatively voluminous docs in an iterative way over time, and with the explicit intent to make them useful to people running *other* Fediverse servers:

> *I do a lot of writing and I built up this sort of mod guideline omnibus. And then when we make policy decisions, often we'll write a blog post framing the question, we open up a discussion among the user base, and then moderators decide on formal policy. And I wanted our docs—like part of this was like reading Darius's work and thinking like, "wow, this is hard-earned expertise that was really formative to how I handle moderation"—I want those same resources to be available to others. And a number of other admins have messaged me and said, I really like [your server]'s documentation & policies, can we adapt these?*

The same admin discussed their process for revising their policies over time by handling minor issues in an ad-hoc way until a pattern emerged, as when multiple reports popped up about a specific issue that was contentious within their community.

> *We want to reason about policy from those specifics. So once we had like three or four reports on [a controversial issue], that's when we actually did the work of opening up a discussion, writing a policy, and announcing the change.*

Minor changes, they said, were fine to just *make*, but "The big stuff, the things that people would get banned for, or that would make a substantive difference in their experience, like Meta federation, those we try to take really carefully." (We get into those more complex kinds of decisions in 4. Complex moderation actions & decisions.) This is in agreement with patterns we heard across most of our interviews—that minor rule changes generally required little to no consultation with larger groups, but that changes that felt more meaningful usually involved deeper discussions, often with a larger group, potentially including the entire membership.

2.3 Initial rule-making process as the first step toward governance

Most teams we spoke with developed their rules, codes of conduct, and other moderation documents in a small group, though a couple of teams began with just one person who made the rules solo. Although we'll be discussing the governance *structure* of servers—BDFL, distributed hierarchical leadership, cooperatives, mixtures of multiple models—in a separate section of this

report, it's worth noting that the initial rule-making process is in many cases a de facto choice of governance structure.

Nathan Schneider of Social.coop highlighted the importance of establishing not only server rules, but also a rulemaking *process*, especially for server teams interested in collaborative or cooperative governance:

> *...that's why I really rushed when the Musk thing happened, building out documentation for Social.coop, such as it is—which is not great, but I just wanted to make it more visible... "Here's the way that you can organize your server. If you want to start a server, if you want to get into this stuff, start thinking democratically from the beginning."*
>
> *My lab built this tool CommunityRule. That's also about, "How can you make it easy and quick to have some rules?" And the idea is not that it's the greatest tool ever, but it's a plea just to say "Get something in place at the beginning so that you have a framework for improving it later. Just get something in place, please, now. Otherwise you'll be stuck with something that all the defaults will just tell you to do something that is, you know, it's just going to be...another weird fiefdom, right?"*

3. Moderation basics

Most of the teams we spoke with who run small-to-medium–sized servers have three to five moderators working at varying levels of engagement. The teams divide their work in various ways: rotations, formal and informal shifts, by natural sleep schedule, and by language and topic. The everyday work of moderation is largely manageable for the teams we spoke with.

> *We have an on-call rotation that's usually one week long. And during that week, we usually have one member who is a go-to person for moderation. So taking action on moderation, being in reports, just writing reports, potentially even on instances whenever a new fascist instance comes up and stays up, or some other clear case.*
>
> *—Social.coop community and tech working group member*

> *We have three people helping with moderation nowadays. Well, technically four—we have one person that's basically doing everything emoji related, and basically only that....we try to send messages through Mastodon to try to coordinate what we think, especially if it's a decision that's a bit controversial. But other than that, it's more whoever takes it first takes a decision.*
>
> *—Masto.donte.com.br admin*

> *I think we're over-provisioned, which is maybe good because it means moderation work is light for most of our mods. We field roughly one to two reports per day.*
>
> *—Woof.group admin*

3.1 Everyday moderation tasks

Simple spam reports came up a lot in our conversations, both in the context of discrete "spam waves" (which affected some servers quite a lot and others barely at all) and in a trickle of routine spammy behavior. As an admin on Tooting.ch, which serves primarily a Swiss regional membership, noted:

> *So what happens, pretty much [with moderation], it's members receive spam. They just report the message and we handle the reports.*

A moderator on SFBA.social, a larger regional server, noted that the bulk of their moderation reports are trivial or non-actionable— either spam reports or reports by an account not on their server, about an account *also* not on their server:

> *...the thing I do most is to suspend accounts, because they're spam. But, you know, that's just super obvious. It's like, "Yes, you're a spammer, goodbye."*

> *...oftentimes the report is coming to us because one of our users was tagged in a thread, and it's somebody on an instance reporting somebody on a different instance, neither of which is related to us. Those are just closed, because there's pretty much nothing we can do about them. Occasionally [in these situations] we'll limit somebody's account if they're posting things that go against our guidelines, especially if they have followers on [our server]. We don't want to just cut them off completely, so we limit instead of suspending...we have once in a while suspended someone's account.*

A moderator on Wandering Shop, which centers on fantasy and science fiction fans and writers, related a similarly tolerable workload for basic moderation actions:

> *We get zero to, you know, if something's going on out there, maybe four or five reports in a typical day, a lot of them are easily dealt with. We have those floods that go on every now and then when a spammer gets loose on mastodon.social, just because we've got to shut down a whole bunch of things just to keep them from nagging our users. But it's not hard. Very little of it has been urgent or problematic material, you know. So we've been lucky in that we don't tend to be a target.*

An admin on Woof.group, a server for the queer leather community, noted that reports from within its membership about behavior (besides spam) on other servers were rarer and often—but not always—required more time and consideration to understand and process:

> *Reports against other instances are less frequent, but more interesting or more demanding. And those come in different flavors. Sometimes it's like, "I don't like that thing and I'm gonna file a mod report about it." And then we just have to look at it and go like, "Well, agreed, what they're doing is maybe distasteful. Does it rise to the level of harassment? Does it require moderator interaction? Should it be us or should it be the remote mods?" Those are tougher questions. And then on occasion, we get easy remote mod problems, which is like, "Somebody sent me an image of a gas chamber saying gays die," and like... easy call, we got your back.*

3.2 Variations across topics and by individual user

For Woof.group, the majority of inbound reports are complaints from members of other servers, often about behavior that is explicitly permitted on their server. According to one admin of that server:

> *I would say probably like 80% come from other servers. And they're typically people complaining about content warning issues....I don't have the numbers in front of me, but I'd guess we act on roughly one in five. The other four out of five are like, "Look, we allow butts here. That's okay."*

Moderators and admins on other servers also noted that certain individual members—and certain topics in particular—attract higher volumes of inbound reports for behaviors that don't break their rules. As one moderator put it, "...we have some folks on the platform who are very intense in terms of moderation needs."

The Gaza-Israel conflict specifically came up repeatedly across interviews as a subject that attracts a high volume of reports, which in turn require both simple and complex decision-making by server teams. An IFTAS advisor researching content moderation told us that:

> *[T]he Fediverse really means different answers for different people—it really is about diversity, even stuff I don't agree with. [...] And then with the Israel-Palestine [issue], it's different because that has boiled over into a lot of malicious reporting. And so it adds a lot of day-to-day stress for moderators.*

Several of the moderators and admins we spoke with mentioned the conflict:

> *Certainly, since last October, I had no idea I'd be spending so much time adjudicating antisemitism. And not just antisemitism, but antisemitism in the context of personal conversations, which are being reported now, which wasn't really something that was a thing maybe two years ago. Like...I don't like what this guy said. So, you know, they open a report to get a user punished, which was not really what was going on, and not what we do.*
>
> *So that's been challenging, because there are really problematic things out there going on. And then there are some others that are like—you know, I'm a little bit hard-nosed about going out and picking a fight on a public timeline, and then running back to me to do something about it. It's not what the moderator's for.*

While a team member on a regional server spoke about the conflict as a source of increased moderation tension and necessary care:

> *The Israeli-Palestinian conflict has been a huge increase in moderation pressure. Both the heatedness of the reports we were getting went way through the roof and the delicacy with which we felt like we needed to approach moderation issues went up.*

This brings us to the next set of moderation decisions—the complicated, messy, and often subtler kind.

4. Complex moderation actions and decisions

> *The question you have to ask is, of course, how much of that time are you "working" and how much are you emotionally fretting over something.*
>
> *—Woof.group admin*

> *Just—when you have to make public statements or make big decisions, for example, the arrival of Threads is and was and still is kind of a big debate. You can receive some public pressure urging you to act and not to act.*
>
> *—Piaille.fr admin*

In nearly every interview with moderators, our interviewees called out special classes of decisions that require more—and sometimes much more—time, attention, energy, and consultation than everyday mod and admin concerns.

Along with contentious Fediverse-wide considerations about things like Threads federation and the Bluesky bridge discussed in detail in Section Four: Federated Diplomacy, we encountered many cases of admins and mods engaging deeply with very fine-grained policy decisions and moderation actions. Those interpersonal conflicts often resulted in intense discussion across moderation teams and also with wider groups including friends, Fediverse admin peers, and—particularly on more democratically governed servers—with subsets or the entirety of the server's membership.

4.1 Vibes and norms

When it comes to moderating the behavior of their own members, the server teams we spoke with vary widely in how much they try to shape behaviors and norms on their servers. Several maintain a light-moderation stance, taking action only on posts that obviously break the server's stated—sometimes largely legally mandated—rules. Other teams, though, make considerable efforts to socialize members toward more harmonious behavior in the community. One admin in the latter camp put it this way:

> *We do a lot of soft guidance. Like I don't want to step in and take aggressive moderator action if a conversation will do. And a lot of times just an email—sometimes a scary one, but often gentle is enough—to be like, "Hey, your interactions with this group here or the way you talk to that server over there have made people uncomfortable. Keep an eye on these things, please, going forward." And either they drift away—they decide that Mastodon isn't for them entirely or the server's not for them—or they manage to get the behavior in check and become more friendly members of the community. Vibes are surprisingly important to a small community and we try not to be overly legalistic about things.*
>
> *For instance, someone might join and pose as a straight, dominant man soliciting gifts from "inferior" gay men. This is a real, consensual fetish for a good number of people. But over time it becomes clear that the account is either a real straight man with very homophobic opinions, or functionally indistinguishable from it. We might step in to have a conversation with the user: we're not here for real homophobia, and if it's play, we need to find a way to make that subtext legible to readers. There's a fine difference, and mods need to be fluent in the subculture to parse it.*

Another admin on a small server takes a similar approach:

> *We also do a bunch of that [moderation] work just by challenging people. Not you know, mean—like, "Hey, this post sucks" type of way, but like..."Why do you think that?"*
>
> *...a lot of times when people are having a terrible opinion online, I feel like one of the reasons people do that is because they think most people will agree with them, and that's why they post it. And then it turns out that they post their terrible opinion, and I have a whole bunch of people who don't think that's a good opinion. So, they don't feel very welcomed, and then they stop sharing that terrible opinion so much. And maybe they never log back on again, and I don't care. But formal moderation...of the local issues is—strangely, I guess, blessedly—it's not actually that bad, once you've done enough work to select people that do share your values up front.*

The same admin expanded on that more high-touch method elsewhere in their interview:

I'm not sure that I've, like, fully actually clicked the suspend account button on somebody, but I have had conversations with people where they ended up clicking the "delete account" button, and that was their choice.

Coaching is…putting your mod hat on and saying "What you did was not cool. Please don't do that again." But in many more words, and with sensitivity to the specific stuff being issued…because basically, I care about all of the people involved… One of the things that helps with having three mods is that we'll have different levels of emotional attachment to the community member involved, and sometimes you want to have a lot of emotional attachment, and sometimes you want to have a more objective third party do the conversation.

The lead administrator of Hachyderm, a tech-focused server run as a project of a non-profit foundation, discussed one of their team's approaches to handling norm-setting and maintenance on the server in ways that actively engage the members who have breached server rules and expectations:

…one of the things that we do, I'm not going to say often, but often enough that I think that it's known about, and we definitely wrote a blog post about it last year, is the freeze pattern that we do. …if you've done something that we feel warrants our attention, we want you to undo that thing. So we notoriously don't delete posts, typically. Because if someone needs a post deleted, their account's going to be frozen and they're going to delete that post as a condition of being unfrozen, right? We try and make it more active so that we're not just this passive kind of cleanup crew running through the instance.

The same administrator noted that the most obviously heated Fediverse issues are rarely the most challenging for their moderators to work through, compared to subtler conflicts between social norms:

… most of the most difficult moderation issues are not the ones that everyone likes to have hot takes about because, funnily enough, easy problems to solve are easy, and you just block people or servers as the case may be.

It's the human conflict stuff that I wanted to make sure [new mods] had a good grip on because sometimes you have people, especially from different backgrounds, where you have them getting into states of genuine conflict, right? You have very American predominantly views of the world coming off of [the server], but that's not necessarily global or correct or exclusive, right? And there are times when they might run into other people's perspectives on anything. And of course we do have users from Europe and so forth on the server as well, and we do have them represented on our moderation team, but just to give an easy example.

4.2 Collaborative decision-making

We asked a lot of questions about how moderators and admins make decisions, either individually or together—or with a larger group potentially including all server members. We touched on one piece of the decision puzzle above, in an administrator's comments about trying to build consequences directly into server rules to make it easier for individual mods to act quickly and independently. Most of the admins and mods we spoke with—including the admin of that previously mentioned regional server with the very specific ruleset—brought up the necessity of consultative decision-making, especially for contentious issues.

A Woof.group admin put it this way:

> *I try to get consensus on anything that is out of the ordinary. So I'll often pose a question to the mod group like, "Hey, what do you guys think about this report? Leaning this way, leaning that way?" And we'll try to talk it through a little bit. And anybody who happens to be available to contribute their expertise or thoughts can come.... I take part in most of the moderation decisions as well. I'm running Woof.group as a BDFL sort of situation, but I try and solicit a ton of input and consensus and will often change my own initial position to align with the mod group.*
>
> *...for big serious questions, things like Meta federation or Bluesky federation, you know, it involves a lot of research: looking at the technical aspects of Mastodon and how federation works, asking what-ifs, consulting with peers, writing up a policy position, soliciting feedback from membership. I can easily burn 40 hours on an issue. And I try to do that maybe five times a year for big stuff. But I think it's worth it. Policy is never going to be 100% popular, but I think we have community buy-in because of our process of writing things up carefully and soliciting feedback.*

An admin on hcommons.social, a project of Knowledge Commons (formerly Humanities Commons) at Michigan State University, spoke about indications that an issue required wider and more substantial discussion:

The more there are questions about reports that we are uncertain how to respond to, we have internal conversations among the team, just to say, "This is what I'm thinking, you know, am I reading this right? Do you see an issue here that I'm not seeing?" So we'll have those conversations.

But where things cross the line over into feeling like we're either going to set a precedent, or this requires some kind of policy. First of all, we developed a Code of Conduct before we launched the instance, to make sure that we had some baseline agreements with the folks who were coming to us, that they were going to adhere to this set of guidelines for their interactions on the network. So we have that to fall back on to say, you know, when something is in violation of those principles, and it's clear.

But where we have cases that don't, that aren't really directly addressed in those guidelines, but feel like we need some sort of community temperature-taking, or some sort of permission from the community to handle in a particular way. We'll go to the community and post as an instance only post saying, "[server name], we got a question, how should we handle this kind of thing?"

A moderator on a cooperative server brought up a series of incidents surrounding reports about a small number of members posting messages repeatedly flagged as misinformation.

...it was quite intensive work, essentially because we had a few members in [the server] who were being reported from the outside.... They tend to be the most interesting cases because, you know, sort of because of our approval process and so on, the vast majority, we don't get a lot of bad actors, clearly bad actors, you know, like spammers or something. It's too much work to sign up, essentially.

So when you do get a report and it's, you know, has substance, it's quite tricky because these are usually, really members of the community. So we ended up having like—and I didn't do a lot of this work, like [moderator name], who is amazing, reaching out to the people and saying, "Okay, so we're getting reports on you, you know, saying things which people consider misinformation."

So this, after discussing...it was like, maybe we need to update the Code of Conduct, but what exactly can we put there? Because if you say, "You cannot deny, you know, vaccines," this would completely alienate these people, right? [The membership was] actively opposed to updating and saying, "We cannot talk about the [subject]," or...whatever keyword you want to say.

In that instance, a large-group discussion resulted in a compromise position that permitted members to post controversial/gray area messages about aspects of the controversial topic, but required the use of content warnings to contextualize them. The moderator we spoke with recalled

that this was acceptable to some people who were posting the messages flagged for misinformation, but not for others, who moved on from the server.

4.3 CSAM and copyright complaints

We discuss legal considerations in detail elsewhere in this report, but it's worth noting that the few interviewees who mentioned copyright or CSAM concerns suggested that those kinds of moderation were a.) special cases that required extra attention and b.) rare. A moderator on SFBA.social, which is a larger server with open registrations, noted:

> *We've had some copyright complaints that were not valid DMCA requests, where we usually use them as opportunities for user education more than we did acting on a takedown. I guess we have had one instance that I know of, of CSAM, where we reported it to the NCMEC.*

Another admin of a smaller server reported that CSAM was the exception to their otherwise relatively autonomous collaborative moderation norms:

> *Like the only thing that I step in for and say, "I'll handle this," is CSAM because there's a legal reporting requirement, and it has to come from the business. Otherwise every moderator has power to respond to everything unilaterally, and we trust that mods make good decisions.*

4.4 Moderator mental health

Notably, no one we spoke with in our core group of server teams reported a high level of moderator stress or burnout, and we attribute this in part to the fact that we necessarily only spoke with teams who had time to set aside for conversations with outside researchers.

For context, in the most recent IFTAS survey in 2023, just under 22% of 129 moderators or admins responding to a question about burnout reported experiencing "burnout or mental health issues due to [their] moderation activity in the past 12 months." (Sixty-nine percent of admins and moderators who responded to the survey were affiliated with servers hosting fewer than 1,000 accounts, and thirty-one percent were involved with servers hosting fewer than 100 accounts, so the IFTAS sample appears to include a lot of servers in the size range we focused on.)

The teams we spoke with tend to have several moderators on staff, to have plentiful (if still insufficient, in their own assessments) documentation for their teams and their members, to have achieved basic financial stability, and to maintain some control over the flow of new member accounts. Although it's not necessarily true that there's a causal relationship between the way these teams run their servers and their relatively non-traumatic experiences as mods and admins, we do suspect that some of these factors have the effect of reducing stress and burnout by reducing workload and easing routine anxieties.

Even so, some teams reflected on the toll of certain kinds of decisions, and the ways in which the less visible emotional aspects of moderating human interactions don't disappear once a moderator or admin steps away from the computer. One admin offered an anecdote to illustrate the effects of moderating interpersonal complexity:

> *...there's a number of times when—like, Christmas Eve with my family, I look at my watch and it's an email from a user that is really upset that a moderation decision is transphobic. And...I take that really seriously. So I have to go anonymize this case and discuss it with like eight trans friends and see what they think about it and make sure that I'm making an ethical decision because I have multiple users accusing each other of harassment and not all of them are necessarily right. You know, that sort of thing really drains you. But thankfully, those cases are infrequent. They're exhausting, they're stressful, but they haven't happened enough to ruin the server moderation experience. ... the problem is like, once it's in your head, it's like, oh God, you can't let it go.*
>
> *I mean, that night on Christmas Eve, I think I didn't sleep at all. I just stayed up thinking about it all night because it was, you know, here's somebody who's suffering and they feel intense pain and they also feel wronged by the moderation decision about that pain. And oh my gosh, I hate to see that person suffer. And that is really difficult to think about.*
>
> *One of the biggest emotional challenges as a moderator is that people can experience severe emotional distress, even self-harm, in response to apparently innocuous behavior. Their report implicitly asks you to judge whether their pain is proportionate and warrants intervention. We try very hard to treat these cases with nuance and empathy.*

4.5 Proactive work to reduce moderation load

An admin of Hachyderm—which maintains open registration—spoke with us about the way their server's moderation load became too heavy for their team to manage easily after the autumn 2022 Twitter migration. While they were also reworking their moderation team and processes, the admin spent a month of free time mapping out—and collaboratively cross-checking—a network of Fediverse servers known for hateful, abusive, and illegal content and actions. Once they'd suspended the resulting list of servers, their moderation work was cut in half:

> *Once we had that done, our moderation burden dropped considerably. So we went from receiving what we were receiving at the time, 20 reports or so a week, maybe more... enough that even with the team that we had at the time, which was double its current size, it was still hard to keep up with. Now we receive...less than 10 reports per week. And most of it is inter-human conflict, which is what we want. Because ideally there'd be no—but you know what I mean? ...the goal of being preventative is for the report count number to drop to zero for a real reason and not just because the report queue is allowed to stagnate.*

The adoption of “worst of the worst” blocklists also came up repeatedly as a way of keeping a handle on moderation workload and protecting server members from needless attacks—for more details on blocklists and the way our interviewees use them, please see 1.6 Shared blocklists and shared blocks.

5. Moderation teams

Most of the server teams we spoke to had a small number (3-5, sometimes a few more) of designated volunteer moderators working informally, sometimes in designated shifts and sometimes in less structured ways. A few servers had a more formal internal structure, including the generalist cooperative server, which has a Community Working Group from which moderators are drawn, and the scholarly server, which receives attention from specialists on the staff of the larger academic project hosting the server, including a community development manager, a user engagement manager, and a UX specialist.

Most moderation teams we spoke with are physically distributed, but some affiliated with regional servers meet up in person both informally and for official meetings of nonprofit entities. Mod teams use a range of systems and tools to communicate with each other, which we discuss in the Tooling section of this report.

5.1 Finding the right people

The teams we spoke with identified many factors in the selection of moderators—without even being asked directly about that process. The server teams that brought up "coverage" or related ideas defined them in different ways, but time zones, language fluency, and cultural fluencies all came up. Several teams also spoke explicitly about the need to find moderators with the right approach, both culturally and in terms of individual orientation/personality, and about the need to identify people they could trust, often by knowing (or meeting) potential moderators offline or otherwise understanding their history and experience.

We'll look briefly at each of these factors, which all—ideally—work together to produce collaborative moderation teams that trust each other and are worthy of members' trust.

With few exceptions, language coverage was important even to servers focused on largely monolingual communities. As Hachyderm admin noted, human linguistic skills remain essential for moderation work:

> *We both try to accommodate time zones and other languages than English ... before we could expand our moderation team, we did have people volunteering their time to help with translations because Google Translate sometimes can't pull just a lot of stuff out of a post. So we would have people that we could rely on for non-English posts...*

Several teams brought up attempts to broaden their moderation teams to include backgrounds and cultural fluencies their original admin/mod teams lack. A Woof.group administrator noted:

> *...then it's time zones and cultural expertise. So I want to make sure that we have—and this is something where the number of axes in which I'm not well qualified to direct and moderate is large—I'm not going to moderate women's issues as well as a woman can. I've done a lot of reading, but I am ultimately a cis man and that's limiting. I'm not gonna moderate Black issues as well as a Black person can. And so I want to bring in lots of moderators with diverse perspectives, but I'm also limited by a small pool of people with experience who've been active for a while, where I can vouch for their character, who can be trusted to make decisions. Moderator action often contributes to inter-instance conflict and collapse; each person on the team brings some risk.*

A member of the moderation team for SFBA.social related that they have had workload challenges and were actively trying to increase their moderation capacity while also building a diverse team:

> *So, yeah, we're trying to grow the moderation team and try to have a lot of diversity in that team. And that's really where I think our most pressing needs were in the past months, just to make sure our moderators aren't burning out and they can take a few days off and we have, you know, enough defense in depth on the team so that if one is going on vacation and the other one is sick, we still have people looking at it.*

A deep concern voiced by several admins is the challenge of finding the right people to moderate, with an awareness that moderator approaches vary widely—and can also matter a lot for the server's members.

Moderators we spoke with called back to their experience moderating on other internet platforms like Discord and IRC—and also to the immense usefulness of having done offline community work before attempting it in the flattened spaces of online community:

> *...people, I think, underestimate the challenge. Like if you become moderator for the first time and you haven't done community management in person.... Having to deal with that stuff in person gives you a certain degree of experience, equanimity, vibes awareness. It's so hard to characterize. But if you start doing this for the first time and you see all these moderation tools like defederation or blocking or whatnot, I think maybe you can assume that it's your only option. But so much of what we do as moderators are subtle discussions with users, or even making posts that are tangentially associated with the topic that's currently under discussion, in a way that calms tensions and creates some community agreement. Sometimes we'll have arguments on the server. And one of the things I've done is make a careful post that acknowledges the argument is happening, and here's why there's difference of opinion, and that's okay.*

A Masto.donte.com.br admin spoke about their decision *not* to expand the server's userbase as a function of the difficulty of building a trusted moderation team, drawing on their early experience

with IRC where members of a channel had interacted positively for months and been accepted as moderators before attempting a takeover of the channel:

> *...it was a conversation that we had with the mod team, which was how many people we think we can actually mod without having a blind spot. And like, yes, we could try to grow the moderation team, but it was also like a question that I still have, because I come from the times of IRC...and I remember having situations where we had like a channel with a couple of friends. And then the channel grew. And then we were talking with the same people every day—like, people that we didn't know personally, but we knew from talking for a very long time and that eventually got promoted to moderators and turned out to like pretending the whole time. And actually try to take over channels like, after being with us for months.*

> *... since I'm not in Brazil anymore, it's a bit hard sometimes to get a gauge of new people that get in the server and don't really know even third parties. [The server's current mod team] all got in together, because they knew each other from Brazil. I didn't know them at the time, but I know some of them now...we met when I went to Brazil, but they knew each other already.*

A Woof.group admin spoke frankly about the risks of bringing in moderators without a solid sense of their history:

> *...one of the failure modes I've observed in other instances is that they selected moderators without really knowing their history. Sometimes those moderators are emotionally volatile or younger, or maybe the mod makes aggressive choices without seeking consensus, or they're not well aligned with the rest of the moderation team, and reaction builds in the userbase and then other mods step in and you wind up with this ugly conflict and it seems to escalate, right? Like instances will implode because of moderator selection. So I'm really cautious about bringing in new moderators. The criteria for me are basically you need to have good vibes and—this is an incredibly subjective position to take, but—you should be able to handle disagreement and difference of opinion without internalizing it. You should have experience in the real world, some sort of grounding out in actual leather play. You need to be able to parse when a post is "You have no idea what you're doing" versus, "Oh, this person definitely plays and it's hot." You can tell that if you've been in the community for a while but for someone who's new, it's not always easy to see. And you also need to be active on the server. There are lots of people who I would love to have as moderators who just don't use Mastodon that much and so they wouldn't be effective. You have to have a finger on the pulse of the group.*

The same admin shared their experience trying to build a moderation team that could collaborate in good faith:

> *One of the things that I try really hard to do is to be comfortable with people doing things that I don't like and to allow other people to persuade me of that being okay, and to have some like good-faith interplay in the mod team. And when I select moderators, I want people who have that same kind of energy. We should be looking to collaborate. We don't necessarily have to agree, but we should be able to come to some kind of defense of the decision.*

5.2 Onboarding and training

Some moderators reported having been given informal orientations, but few moderators reported any formal training as part of their introduction to their server's team, though the amount of documentation provided varied widely. One moderator's experience was especially hands-off:

> *...I kind of looked at all the previous moderation decisions that they had made. And I asked a lot of questions. And that was basically how I got trained.*

Other mods noted that they'd joined the moderation team during a period of rapid change or high stress—which, we suspect, is when a lot of moderation teams expand—and therefore received little or not orientation, though one mod with this experience mentioned that the team's documentation had subsequently improved.

On the other end of the continuum, the Hachyderm admin we spoke with noted that they onboarded all their moderators themselves:

> *For the moderation, for the culture, I onboarded all the moderators. Personally, I went through and made sure everybody understood...*

And this same admin also reported using internal, non-public documentation to guide moderators through their work, both to clarify processes and to help new moderators understand how to shape their own experiences as mods in ways that preserved their mental health:

> *...we do have internal documentation about how to identify, how to meet, how to discuss, how to opt into depending on what, because everyone has their sources of trauma too. We don't want anybody to get a face full of—you know, whatever. There's a lot of violent and illegal and sometimes both stuff out there*

Notably, very few of the server teams we spoke with indicated that they maintain internal documentation for their moderators—presumably because most moderator teams are so small—but we think that internal docs, along with careful onboarding, are probably a good way for newer or expanding servers to support their moderator teams.

6. Additional moderation resources

Moderation resources written or mentioned by server teams we spoke with, in addition to the Mastodon/Hometown pages and off-site moderation documentation maintained by each server team and linked in 2.1 Documentation types and links above.

- IFTAS Connect, a community for server teams from Independent Federated Trust & Safety
- IFTAS Moderation Handbook
- Run Your Own Social, a guide by Hometown maintainer and co-author of this report Darius Kazemi but mentioned by another admin we interviewed
- Three Gates of Speech notes on a wiki run by one of the owners of Fediverse server Merveilles.town

Section Three: Server Leadership

Introduction

In the previous section, we introduced underlying factors that shape governance in the Fediverse and investigated current approaches to moderation, or the governance of members and content. In this section, we'll discuss the **governance of Fediverse microblogging servers and server teams**, including how decisions are made, how authority and responsibility flow, and how infrastructure (both technical resources and human time and attention) is chosen, allocated, managed, and sustained. We call this layer **server governance.**

These layers of governance could also be—at least in theory!—distinct from the kinds of **legal entities** server teams form and inhabit, but in practice, we've found that formal legal structure and governance models are often closely connected. These are discussed throughout this document where applicable.

The independence afforded by the federated model of social media allows for local experiments in the governance of servers themselves. The majority of servers we're aware of on the Fediverse run along extremely informal and top-down lines—most obviously in the case of single-person servers, but also most small and medium-sized servers, and even most of the unusually large servers. Because we're interested in structures of server governance that extend beyond or rework these cultural defaults, we intentionally selected a range of governance approaches in our research sample.

We spoke with members of 11 teams who operate their servers in various ways:

- top-down, BDFL (Benevolent Dictator for Life)/BDFN (Benevolent Dictator for Now) structures with consultation among moderation/admin team members and varying degrees of consultation with server members (5)
- projects of not-for-profit entities including French and Swiss non-profit organizations, a technology foundation, and a university lab, with relationships to their members ranging from the aspirationally democratic to the strongly consultative (4)
- formal cooperatives (2)

Interestingly, very few people we spoke with considered their governance structures to be fully settled and aligned with their collective sense of the best way to run a Fediverse server. This was true for both very informally run servers and those with many layers of process, bylaws, and documented rules. An interviewee working toward finding the right structure for the server they help run even pointed out the real rarity of any formal structure on Fediverse servers:

> *I actually have been working on going through... the FediDB to go through and be like, what is the organization for all the top X servers on here to see? And, you know, most of them is "none"! And there's a handful of for-profit companies and a very small number of something else, has been my sense.*
>
> *—a legal advisor to a larger server*

This sentiment was echoed by an advisor to IFTAS:

> *I'm not really seeing a lot of different experimentation. I'm seeing people aren't aren't particularly comfortable with the basically autocratic, the benevolent despot model, which I'd say...most people know it's problematic, but they don't, other than the co-op model...have good alternatives.*
>
> *—an IFTAS advisor*

In contrast to moderation processes and norms, which are extensively developed across many servers, server governance on the Fediverse beyond informal and autocratic defaults is still nascent, and resources for server admins interested in trying alternate structures—especially resources including detailed and adaptable examples—are thin on the ground. (We'll echo this finding in Section Four: Federated Diplomacy)

We think there's a lot of room in today's Fediverse for projects focused on expanding these kinds of resources and building community and connections between server operators interested in trying out more structured, more participatory, and more democratic forms of server governance. As a first step in that direction, we're using this section to document the models and structures of server governance that we encountered in our research.

Key observations

- **Especially in terms of server and institutional governance, it's still very early days on the Fediverse.** We intentionally spoke with server teams who'd given governance careful thought, knowing that they represent the far end of the Fediverse governance continuum, and still heard that many of them don't believe their governance models are fully thought through or fully implemented.

- **More teams aspire to participatory or democratic governance than have the resources to implement it.** Most of the server admins and teams we spoke with aspire to democratic forms of governance, but the work of figuring out exactly how to do that is a substantial barrier for admins who operate Fediverse servers as a sideline to their other work. And beyond that barrier, identifying and implementing models of participatory governance that allow for skillful, sensitive, and rapid decision-making is a challenge even for experienced co-op leaders and members

- **The first few decisions server operators make have disproportionately large effects on how the server will be run.** Early decisions about how a server will run—how rules will be made, how power and accountability will flow, what the operators will ask of members, and what software will be in play—have strong shaping effects on future governance, and those effects are harder to overcome later in a server's lifespan; we think it's a good idea to consider these elements as early as possible in a new server's life.

- **The Fediverse microblogging toolchain supports little variation in governance.** The affordances of Mastodon (and the Hometown fork) and other infrastructure required to operate a Fediverse microblogging service support top-down decision-making; other governance models require tinkering and additional software.

- **Server stability is hindered by lack of space for succession planning.** Succession (and end-of-server-life) planning is a subject many server teams note that they need to think about, but which doesn't have a lot of obvious precedent and isn't top of mind for teams that spend most of their available time/resources on essential maintenance and reactive (non-derogatory) decision-making.

- **Server members aren't universally invested in intense participatory governance.** The appetite of server members for participation in server governance varies widely, but only a minority of members of even the most participatory server we engaged with actively participate in decision-making and the discussion that supports it; we think it's wise for server teams interested in more democratic governance to consider participation options of varying depths.

1. Three models of server governance

1.1 Independent top-down governance

Just over half the servers we looked at are governed in a largely top-down or explicitly BDFL/BDFN way by a small team without formal oversight by an organization or board, and often with a founding or early administrator setting the overall direction and culture. This 50% proportion actually under-represents the prevalence of this model and is non-representative of the Fediverse as a whole: an overwhelming majority of Fediverse servers we're aware of are run by single individuals or small groups, sometimes in informal consultation with server members over high-profile issues.

The main benefits of this model we heard:

- Simplicity and speed of setup—it's possible to run a top-down Fediverse microblogging server using just core software and a group chat for moderators to communicate with each other privately.
- The ability to maximize the cultural benefits of having a founding admin or small group with exceptionally strong community management skills. This point is particularly apparent in discussions with operators of small servers that intend to stay small, and with the admins of servers that focus on a narrowly defined community, like members of a subculture or people within a metro area who share a specific political orientation.
- The ability to run a server without asking much of its members. ("Nobody wants to do the legwork of becoming a co-op or doing any additional hassle. It's like, I ask for mods to join and I get maybe one person if I'm extremely lucky. ...I don't see a need to do stuff that nobody's asking for.")

The main downsides we heard about:

- A gap between democratic ideals and a sense of what the team's had the resources to implement, or what works best in practice. (We heard about this across multiple models of governance.)
- The sense that larger servers in particular run up against resource and organizational constraints that are difficult to manage without building more formal or complex models of governance, including difficulty staffing moderation and admin teams at sustainable levels and challenging financials. ("I think we're looking for a model that lets us handle turnover in the administrative staff and some model for sustainability. We obviously don't have the type of funding resources where you could have permanent staff, but it would be nice to have a board, where the board can help find people to step into various volunteer roles...")
- Several members of top-down teams we spoke with flagged sustainability (financial and human) and server longevity as challenges that they grappled with, and highlighted the

> need for baseline financial support from server members and the establishment of a team of trusted, collaborative colleagues to make it possible for everyone involved in a server's operation to take time off, weather illnesses and crises, and potentially resign their duties in the future without taking the whole community offline.

The actual legal structures underpinning the top-down models of governance can vary widely. Some of these servers have no legal entity tied to them and are run by private individuals. One such server is run by an individual with a job that requires them to not handle money; this server operator partnered with another person whose personal bank account holds the actual (small) funds required to keep the server running month to month. Other top-down servers have more formal structure. An admin of Woof.group tells us

> *we actually incorporated last year. So after the massive influx of users, it's like, we need a little bit more legal protection. We need some sort of independent structure and funding. So Woof.group is now a self-sustaining-ish LLC. We have lawyers who we pay real money to. And they give us real advice on issues like CSAM and help write our terms of service agreement.*

Interestingly, SFBA.social is likely moving from being a BDFL/N with a non-profit fiscal sponsor to being a BDFL/N with a standard LLC similar to Woof.group's current structure. This became necessary when the server's nonprofit fiscal sponsor, Open Collective Foundation, announced their sudden dissolution in early 2024. SFBA.social says they are likely to move to an LLC due to trouble finding a replacement fiscal sponsor as well as the cost of forming their own nonprofit entity:

> *The quote we got for incorporating as a nonprofit was $7,000 to $10,000. And I was like, Oh, never mind. Like that's way out of our budget. Like we don't, we make that [in] a year.*

Under their prior fiscal sponsorship arrangement, their ability to act as a non-profit was critical to their funding. They explained to us that,

> *being able to do nonprofit things is important for us. Cause it's actually a pretty substantial portion of our funds are matching funds from—some of the folks on our team work for tech companies that will match time and funds for their stuff.... And those checks kind of come in very slowly. When they do, they're usually pretty big.*

Matching funds will no longer be an option for this server if they lose their fiscal sponsorship or fail to incorporate as a non-profit themselves.

1.2 Cooperative governance

Two server teams we spoke with, including one "core" server we did multiple interviews with, and one newer server we engaged with more briefly, run as formal cooperatives. Notably, several other server teams expressed interest in learning more about cooperative models, and especially about the details of getting from "zero to something" in the process of establishing a co-op.

Even in our tiny sample—we spoke with only two servers run as formal cooperatives, and there aren't very many active across the Fediverse—we encountered two significantly different approaches to the model: the larger and more established co-op runs a very full-participation model (one admin compared the server to the Park Slope Food Co-op, which is renowned for requiring all members to contribute labor), while the newer server emphasizes governance by its board and working group leads, with member consultation on critical issues and open meetings.

The main benefits of this model we heard:

- The chance to work in a full-throated way toward new and—in the optimistic view—better, ways for people to be together on the internet, with participatory decision-making and communal support (both financial and through time spent working on the server community) built into the server from its foundations up.
- The potential for the kind of long-term stability that eludes many Fediverse servers whose governance models rely on the availability and interest of a single lead administrator or a small team of operators. (One co-op founder told us: "I think organizational resilience and stability, with particular view to the financial side, is key—you know, many, many instances are running on Patreon, which is okay-ish… But I think if you're going to be decentralized, then you're going to need actual careful thought given to organization design and finance. And that's why I'm an enthusiastic co-op-based instance evangelist.") This factor also applies to some of the models we'll discuss in 1.3 Non-profit entities as a middle path.
- The chance to practice democratic decision-making and governance online as a way of (re)building these skills and normalizing participatory practices and expectations across a populace in ways that could, ideally, seed stronger civic/community participation both online and offline.

The main downsides of this model we heard:

- It's hard for very small servers and those run in their operators' spare time to work toward full cooperative status—the financial, legal, and especially social considerations require a more work and time than many individual admins can spare, and it's not clear to many people where to begin and how to succeed in gaining critical mass.
- Cooperatives face headwinds in the form of the (to use Nathan Schneider's phrase) "implicit feudalism" present in the default settings of many open source systems, including popular

Fediverse microblogging software; the cultural and organizational complexity involved in setting up and running consensus- and discussion-based governance processes; and the complexities of establishing legal cooperatives.

- It's tricky even for full-on cooperatives to find the right balance(s) between participatory ideals and server members' varied and fluctuating interest in and availability for deep engagement in communal self-governance.

The admin of a regional server noted that they'd worked with non-profit associations in Europe, and they were interested in cooperative structures, but that the process of getting started and building a core group to move forward on their existing server—which has been running for several years—was really challenging:

> *...the hardest thing is the financial-slash-organizational thing. I can probably get in touch with the lawyer in Brazil, figure that out. And I'm pretty sure it will take, like, a little bit of time because bureaucracy, but, that's not the hardest part. I think the hardest part is actually finding the people that at least want to kick off the thing and be involved...and setting up the guidelines on how this will work and all of that stuff.*
>
> *There is a part of finding some people...at least two people to be actually involved in managing the server.... I did a little informal, like, "Who would be interested?" and three people in the server [were] like, "Yeah, yeah, I would, I don't know exactly doing what, but I would!" And I'm, like, yeah, but three people on a server of a hundred active users might not be enough to keep the thing going for a longer period of time. Maybe we need a bit more or maybe we start with those people... and then we set up a more permanent structure and more clear roles.*

Of the two formal cooperatives we spoke with, CoSocial Community Cooperative is a Canadian community service cooperative under British Columbia law, a special designation that has been described as a cooperative that exists for the benefit of its community rather than the benefit of its members. CoSocial sustains itself via membership fees. According to a founder of CoSocial:

> *It costs $50 a year for membership fees. People have the option, if they're feeling generous, to put in $250 instead of 50 for a supporting member. And a few do. And then we have...organizational membership, which is also $250 a year. The other thing is that we've been doing this for so long, and if we do continue to grow and get to a few thousand members, we're either going to have to lower our prices or figure out a way to share money, because, you know, 50 bucks a year is plenty. I mean, that's way more than it costs to actually provide a Mastodon account. So, but, you know, all of this is really arm-wavy, because it depends on us proving that we can last and grow.*

This participant is optimistic that a paid membership model will result in an easily self-sustaining cooperative. They also explained to us that a future possibility for CoSocial could be forming a subsidiary non-profit that could allow them to apply for grants and/or solicit organizations and individuals for tax-deductible donations.

The other formal cooperative, Social.coop, is not a non-profit itself but operates with the assistance of a fiscal sponsor organization, itself a UK cooperative. The day to day operation of Social.coop is carried out via working groups. Its legal and financial work is carried out via one of these working groups. Expenses are proposed, discussed, and approved asynchronously in a Loomio forum, and the actual disbursement happens via Open Collective (the software, not the foundation mentioned elsewhere in this report) . This is discussed in more depth in 5.1 Cooperative decision-making.

1.3 Non-profit entities as a middle path

We spoke with members of server teams that are each projects of a Swiss non-profit association ("Association à but non lucratif"), a French non-profit association ("Association Loi 1901"), a US non-profit foundation run as a cooperative, and a nonprofit commons network run by a research lab at a US public university. (We've put the server affiliated with the non-profit cooperative in this category rather than the "Cooperative governance" category because although the foundation runs as a co-op, the Fediverse server itself is run along top-down lines.)

Each of these servers offers a model for institution-building on and around Fediverse servers in ways that differ from a fully co-operative model but still involve some degree of participatory (and transparent) governance.

- Both European servers that run under formal non-profit associations hold required meetings of their general membership and consult with members to varying degrees on certain issues.
- The server that runs as part of an academic commons network maintains a highly consultative relationship with its members that one advisory council member traced to a faculty governance model inherited from (perhaps a previous age of) higher education.
- The server that runs as a project of a technology foundation is governed in a top-down way by a group of infrastructure administrators and moderators, though the foundation itself—which was established in 2023 and is still being built out—is being developed to run as a cooperative.

Upsides we heard about:

- Although it doesn't prevent unexpected dissolution, the formation of a legal entity or institution tends to clarify accountability, reveal financial situations, and signal that sustainability and longevity are priorities for server operators, potentially increasing both trust and trustworthiness.
- Nonprofit entities encourage—and to some degree, require—clear decisions about organizational design and governance, including thinking through bylaws and board formation, as well as regular reporting on donations and spending. In the case of the server

embedded in a university structure, they have access to legal advice through the university's Office of General Counsel.

- Pathways to participatory governance, whether of a parent entity or a Fediverse server, provide some of the benefits formal cooperatives confer, like practicing the skills of democracy and engaging server members in the complexities of reconfiguring the social internet, without requiring as complete a commitment to co-op models.

Downsides we heard about:

- In some jurisdictions, incorporation as a non-profit organization is complex and very expensive—one US-based server in the unusual position of already having volunteer legal help was quoted $7,000–10,000 for the process of establishing a non-profit entity, which is more than the amount the server receives in donations in a year.
- Acquiring a fiscal sponsor has been a good option for many US-based entities, but the dissolution of the Open Collective Foundation has been very destructive for Fediverse servers, who are having a very difficult time finding replacement sponsor organizations.
- Working with boards, running general membership meetings, and doing legal and financial compliance work eats a lot of time and energy, which can feel out of reach for small teams.
- One member of the server affiliated with the lab at the US public university expressed concern that contentious political issues around speech at US universities could bleed over into the Mastodon server they provide the infrastructure for. It is unclear to them whether speech on the server reflects on the university. They also pointed out that US public universities have responsibilities around FOIA and other US transparency laws that a private operator does not.

A server admin spoke frankly about the gap between their beliefs about how things should be run and the risks and vulnerabilities of formalizing out of a BDFN model and into a more democratic or board-oversight model:

...my own bias would be like, "everything should be democratic". And we should have elections and a board... One of the most important pieces of advice I ever got from a dear friend who does organizing work for queer nonprofits was like, "Do not get a board. As long as you possibly can avoid it. Because you'll have to deal with things like "your new diversity and equity head on the board, who just won their election by a landslide, turns out to be saying a lot of racist things on Twitter. And now is in charge of approving the training budget for themselves for remediation."

You get people who are really interested in power and anarchism and democracy for sort of... formalism's sake. I believe firmly that structure is important, but I'm more interested in "How do we keep people's emotions healthy—acknowledge their struggles and diffuse tensions and produce a community which is healthy overall?" The really messy, anguishing work at the edges. The institutions that I've seen work really well—sometimes for decades—often they have a core group or one person who really sets the tone. And that has intrinsic scalability limits, right?

2. Specific structures and patterns

2.1 Membership discussions and meetings

Whether through formal meetings, discussion boards, proposal-and-voting systems, or informal calls for feedback and discussion within Mastodon or Hometown, most (but not all) the server teams we spoke with engage their members to varying degrees in decision-making and governance. Several server teams report having engaged in discussions with early members and potential members during initial server setup, but the majority brought these aspects of participatory governance online after the server was up and running.

One co-op server has an annual general meeting as required by law in the jurisdiction where they are incorporated. In this meeting, finances must be reported to the membership, elections of officers and board members are discussed, along with other agenda items required by law.

2.2 Boards

A few server teams we spoke with have boards—one co-op server governs itself via a board and working groups, and most servers affiliated with non-profit or academic entities have contact with boards at the entity level, and board membership often overlaps with server leadership.

2.3 Working groups

Both cooperative servers we spoke with are organized entirely or in part around topical working groups: one has five working groups organized around Community, Finance, Legal, and Tech, with a new Organizing Circle working across these groups. The other co-op server has working groups organized around Communications, Finance, Membership and Outreach, Technical Operations, and Trust and Safety.

On the larger co-op server, which is intensely participatory in character, the working groups make "operational" decisions that flow from "strategic" decisions made by the membership via proposals and voting; on the smaller co-op server, working group leads make decisions for the server, in collaboration with the board.

2.4 User advisory groups

The academic-affiliated server's parent project (a non-profit commons) recently established a user advisory group, which draws its members from the users of any of the project's initiatives and services and functions as a user-research/focus group of, in one member's terms "super users." Although this particular group will be focused on the Fediverse server only some of the time, the model itself could be easily adapted to work for other server teams focused only on server operation.

On their blog, the project's leaders note that they established the group to:

- Empower our users to have a larger say in the development of the Commons
- Create opportunities for our users to connect with our team and within the Commons community
- Communicate directly with users whose values align with those of the Commons
- Provide space for the open exchange of knowledge and ideas between the Commons team and our users

3. Paths for exploration

In our conversations about governance structures and models, three cross-model concepts rose up that would, we think, reward further research and discussion.

All three concepts seek to address instabilities in the human infrastructure of the Fediverse: The first is a proposal to more intentionally connect less-technical people with strong community skills with more-technical people interested in running the technical infrastructure of Fediverse servers to make servers more culturally attractive and resilient. The second looks at potential benefits of getting more institutions onto the Fediverse, both by integrating existing ones and building new ones, with an eye to increasing the number of stable entities on the network. The third suggests a line of inquiry into ways of thinking through and building out pathways to greater participation in self-governance for diverse levels of interest and availability. We think any or all of these could serve as a strong backbone for additional research and collaborative building on the Fediverse.

3.1 Connecting people-people and tech people

Many people on teams we spoke with talked about the critical importance of recruiting moderators and co-administrators who were experienced in the challenges and realities of offline and online community work. Members of three different teams brought up the importance of meeting or knowing one another IRL. A founding member of a Fediverse cooperative extended these ideas further:

> *...the thing I've also gotten really passionate about is supporting people...who are naturally community people, and who are community builders. One dream of mine that I'm trying to get funded was an incubator for founders of Fediverse communities [to] draw people in who love connecting people and making spaces fun for other humans, and then giving them the tech support so that tech is not something they have to worry about.*
>
> *To me, one of the deep problems with the Mastodon world and the Fediverse is that it indexes so much on tech skills ... it's a lot easier to find someone who can write code than someone who can make a new user feel really welcome. ...what normal, well funded organizations do is they have really skilled people-people and they have really skilled tech people, and they put them in one organization, because they have money, they can hire them—it's not rocket science! But it is tricky in a context where you have such an underfunded ecosystem, comparatively.*

We think bringing skillful community practitioners into collaborative relationships with skillful tech practitioners to run Fediverse servers is a move that would make the experience of participating on the Fediverse better and richer while increasing the resilience of servers. When this happens now—as it clearly does within many of the teams we spoke with—it's largely a because of the community experience and personal or professional networks server founders can bring to bear on their Fediverse work.

We think there are opportunities for a lot more of these intentional collaborations between people with divergent skillsets, and we *suspect* that many of those opportunities will spring from teams who take structure and sustainability seriously, and from people and organizations who commit to building stable institutions on the Fediverse.

3.2 Easing institutions into the Fediverse

> *I'm pretty sure that if the whole Fediverse continues to survive and grow, we're going to see a ton of institutional instances for universities and departments and businesses and professional associations and things like that, which obviously are going to have fewer sustainability problems, because who owns it and who's responsible for it will be clear, and it'll be a line item in the cost budget, and that'll be clear—and assuming there's a benefit, people will see that as something just like operating their email, right? You gotta have email, you gotta have Fediverse. So I wouldn't be surprised if down the road that becomes a very high chunk of all Fediverse activity, organizationally operated servers.*
>
> *— A founder of a cooperative server*

There's no Fediverse consensus about whether it's an intrinsically good thing to see new entities of any given character spring up on the Fediverse, but we think it's noteworthy that the handful of existing extra-Fediverse entities that have established their own servers or integrated with the ActivityPub ecosystem—*The Texas Monthly*, *ProPublica*, RestOfWorld, Knowledge Commons (a participant in our research), Medium, Flipboard, WordPress, and Threads come to mind—have been enthusiastically embraced by many Fediverse users. Threads has also been exceptionally controversial because of widespread reservations about its moderation practices and its parent company, Meta.

The Threads integration came up in our conversations with nearly every team we spoke with as an issue that had stress-tested consultative, decision-making, and communication processes, and our research sample included nearly every possible choice in relation to federating with Threads. It seems clear that an institution's reception in the Fediverse will be based at first and in large part on their reputation *outside* the Fediverse.

We think the potential benefits of participation by (subjectively benevolent) institutions in the Fediverse are benefits to the commons: If more institutions can offer financially sustainable, appropriately staffed servers and services, Fediverse users gain access to broader sources of information, more connection with people and entities they value, and potentially to servers that provide stable, long-term community hubs for people seeking accounts less likely to be subject to arbitrary shutdowns or mass defederations.

One of the server teams we spoke with, from Knowledge Commons (formerly Humanities Commons), serves as a proof of concept for the provision of Fediverse services by academic-affiliated institutions. As an academic project designed to provide infrastructure—a

repository, blogs, and now a Fediverse server, among other services—to scholars and others interested in commons, Knowledge Commons has an institutional rationale to provide Fediverse infrastructure to members of the public, rather than only to their own faculty, staff, and students, and has found that the majority of their Fediverse members have become aware of Knowledge Commons via hcommons.social, rather than the other way around:

> *...we were approaching it as an experiment and weren't really sure how it was going to go and whether it was going to survive. Quite honestly...it's doing extremely well. And it's starting to influence the ways that we're thinking about the core of the network now.*
>
> *— a founding administrator of hcommons.social*

A member of that server's user advisory council noted their belief that many Fediverse services deny accounts to institutions and organizations on the principle that institutions should be acting as infrastructure providers for at least their own accounts and staff, and potentially for others, rather than taking advantage of largely volunteer labor.

The technical foundation we spoke with, the Nivenly Foundation, was actually established as a home for its founders' Fediverse server, and now serves as a home for other open source projects:

> *... once we realized [the server had] surged large enough that even with the lulls, it was going to need some sort of mechanism to handle financing the server and some other things. We created Nivenly around that and then we decided to wrap other projects in Nivenly as well, so Nivenly is intended to be able to exist without Hachyderm—however, as its first and largest project, it very much is for Hachyderm.*
>
> *—the Nivenly Foundation's executive director*

We also think the benefits offered by at least partly Fediverse-centric new institutions like IFTAS are obvious and substantial, and can help fill many of the gaps—technical and otherwise—identified by the admins and moderators we spoke with, even when the direct provision of Fediverse servers isn't any part of their mission.

Some of these institutions—both formal and informal—serve as gathering places for the meta-community of Fediverse server operators. Many of the server teams we spoke with noted that they participate in off-Fediverse forums or chat rooms designed to bring together Fediverse administrators and/or moderators to share information and provide peer support, including the official Mastodon Discord (accessible to Patreon supporters of the Mastodon project) and other Slacks, Discords, and forums.

The administrator of one moderator forum spoke about the governance challenges arising from bringing together a heterogeneous group of people who are themselves attempting to set and enforce policy for other heterogeneous communities:

> *I've seen moderator groups implode from lack of defined governance. And not around [Threads federation and the Israel/Gaza conflict], they imploded before these came along, but we'll only see more of that where folks are, are congregating in a Discord or whatever and feel they've got it until a real juicy problem shows up and people realize that they might have ideological differences…that lack of structure and that lack of process hurts a lot.*
>
> *[…]*
>
> *[Threads federation has] certainly been a very heady topic in the moderator chat rooms, people endeavoring to be right or be proven right or prove someone else wrong. We have very strict community participation guidelines and it kind of squelches most of that conversation. And that's intended, we're all about—we know you have differences. There's 27,000 service providers. You clearly are coming at this from a broad spectrum of philosophy and goals. We don't really care where you differ, we care where shared practices can help each other and where you're willing to leverage agreement. So, disagree, but don't do it disagreeably.*

One founding admin of a server for French speakers spoke about their desire to gather fellow admins for Francophone servers into a non-profit entity that could coordinate shared blocklists, moderation decisions, and positions; when we spoke, this admin had got as far as holding a meeting with about ten other Francophone server teams, and hoped to be able to devote more resources to the development of a regional meta-federation.

An advisor to IFTAS spoke with us about the potential for connections between servers and teams that might serve as formal or informal meta-institutions:

> *And from a governance perspective, from across server things, I really feel like there's this missing level of between the federation of everything, everybody, and the individual instance, there's this sort of collection level, in Run Your Own Social, you referred to it as a "neighborhood," Darius. And yeah, there's Kat [Marchán] calls them the "caracoles," `@ophiocephalic`'s "fedifams," or the bubbles that are emerging, all in this. And I think that's another place for, for very interesting approaches to governance. It's like, the two most worked out things I saw have both come from an anarchist perspective, which is interesting, because it's a chance for radical democracy with some structure to it.*

3.3 Making pathways to greater participation

A founding member of a cooperative we spoke with the most noted that not all their active monthly users are currently registered as users on their decision-making platform, and that only a subset of everyone registered on the platform tends to read relevant posts—and then a much smaller fraction actually participates regularly in member discussions and votes, with only a few dozen people routinely participating in synchronous gatherings.

The member we spoke with pondered whether that level of participation constitutes success—and if not, what could be done to shift ratios:

> *...the question is, is that good? You know, maybe it's totally fine to have a kind of oligarchic structure like this. But I would like to see a much clearer pathway for people to participate. One thing that we're doing with the organizing working group, or the organizing circle is there's a random component to it. So people can be selected randomly to join, in addition to working group members. I think that kind of thing is really appropriate in this context, just to just to pull people in and give people who might not know how to step up a way...*
>
> *[...]*
>
> *...we built a co-op on the model of a more volunteer-driven participatory, you know, "everybody all hands on deck" kind of approach, but we've expanded to a size where that's not really a reasonable expectation—and we don't have the onboarding to enable people. And so, you know, I think it's something to work on. ...what we want to be is something that's really giving a co-op governance experience to everybody, but we haven't built up the structures to really follow through on that.*

The same interviewee noted that varying levels of participation aren't necessarily a bad thing for all cooperatives:

> *This is one thing I'm kind of fighting in co-op governance right now is this idea that low participation rates are necessarily like a terrible sign. I think it would be very reasonable, for instance, to design, say, a Mastodon instance that's set up so that everything is run by paid staff, but there's one or two things a year where members can participate and have a voice and shape the future of the thing. And it's meant for people who just do not want to be thinking about their cooperative all day, they just want to use it. I think that's totally fine, too.*

We heard similar concerns about a lack of obvious pathways to broadly accessible levels of participation from many server teams, and the other cooperative we spoke with has adopted a less hands-on model for its own server, as noted in 1.2 Cooperative governance. Hearing it from the most hands-on cooperative on our list was especially striking, and we think this topic would benefit from wider discussion and collaborative service-design work that pulls expertise from within and outside of the Fediverse.

4. Governance Resources

Governance resources written or mentioned by server teams we spoke with

CommunityRule (a "governance toolkit" for communities)
Cooperative identity, values & principles | ICA
Social.coop Bylaws
Social.coop Community Working Group Ops Team
Social.coop's How To Make the Fediverse Your Own

Section Four: Federated Diplomacy

Introduction

In the previous two sections, we discussed **moderation** (governance of members and content) and **server governance** (governance of server infrastructure and teams). This section focuses on the third—and overall least developed— of the three layers of Fediverse governance we considered: the governance of relationships between servers.

Two of the most prominent differences between Fediverse governance and central platform governance are that **1.) all direct governance on the Fediverse is local**, and, therefore, that **2.) all governance extending beyond a server's bounds functions via diplomacy**. In the Fediverse, there are no technical means for one server's operators to force another server's operators to take a given action—the threat of defederation (limiting or suspension) by one server or a coalition servers is the *only* built-in lever in the Fediverse for the cross-instance exercise of power.

The power structures of the external world still apply: we heard about server operators forced to take specific actions—or leave the Fediverse entirely—because of legal changes in various jurisdictions, because of both good- and bad-faith reporting of their breaches of local law, and because they were doxxed and their offline lives were affected by it. But within the Fediverse as a system, all inter-server actions turn on diplomacy, ranging from diplomatic-as-in-tactful communications to coalitional pressure campaigns to outright belligerence.

These aren't novel observations. (De)federation dynamics are at the heart of many of the Fediverse's most heated discussions and controversies, but there's a substantial gap between the collective awareness of the complexity of these decisions (see 4. Complex moderation actions & decisions for discussion) and the lack of clear, public policies that might ease and guide them, beyond loose affiliation with the Mastodon Server Covenant.

Key observations

- **Inter-server diplomacy (or, most commonly, cessation of diplomatic relations) is both crucial and challenging for thoughtfully governed servers:** (De)federation (limiting and silencing of accounts and servers) and other inter-server questions are very prominent in administrators' and moderators' accounts of their experiences on the Fediverse, and account for many of the most stressful decisions server operators make.
- **Specific and controversial questions in federated diplomacy have drawn increased attention to this aspect of governance:** Federation with Meta's Threads social network site has served as a stress-test for many server teams, and has nudged many server teams to

communicate publicly about not only their decision, but also their rationale—and in some cases, to make more consultative decisions. We heard much the same but to a lesser degree about the question of opening connections with the Bluesky decentralized network via a cross-network bridge.

- **Server teams' positions on questions of federated diplomacy are more frequently offered as one-off statements than part of a clearly documented set of diplomatic policies.** Policy and process documentation about inter-server governance lags far behind local moderation policies and processes. We think this points to an opportunity to clarify dynamics that are important to both server teams and Fediverse members.

The diplomatic layer of governance is largely undocumented

This gap is especially apparent when we compare federation policies and processes with the relatively rich set of public moderation policies and processes that apply to a server's own members: aside from public lists of defederated servers, most servers don't publish any rules, policies, or norms about their (de)federation processes.

In the simplest terms, this means that server administrators and moderators make decisions about whether/when to limit or suspend federation with other servers and with individual members of other servers—but the criteria for their decisions are often unclear, and sometimes inconsistent. *This is especially detrimental to would-be server members trying to sort out which Fediverse server to choose*, since defederation has strong effects on the way a server's members will experience the Fediverse, including how much abuse, harassment, hateful or violence-inciting speech, and spam they're likely to see.

To be clear, "Just add documentation" isn't a magic cure for a lack of clarity or a surfeit of complexity, but we think the process of composing, validating, and publishing federation and defederation policies can be a helpful forcing function for working out what the underlying principles actually are—especially when the policies are specific and thoughtfully customized to suit a server's aims and character.

Policy and structure vs. technical tools

In this section, we'll look at (de)federation decision-making and policies; for more tech-focused information on shared blocklists and similar tools, please see 1.6 Shared blocklists and shared blocks. (The governance of *shared blocklists themselves* would be a worthy subject for future research, but was beyond the scope of our project.)

1. Federation as remote moderation

Some server teams moderate members of other servers ("remote users" in Mastodon's documentation) largely as though they were members of their own servers ("local users"), but many built-in local moderation tools like warning messages and account freezing aren't available for remote users—and, of course, those remote users also haven't agreed to and often aren't even aware of the moderating server's rules and principles, so it's not a tidy parallel.

Additionally, some server teams who maintain relatively restrictive rules and norms for their own server's members—around things like nudity, content warnings, what's considered off-topic or an inappropriate use of an account, or unacceptable rudeness—but lack clear policy about whether or when they limit or suspend remote users for breaking those same rules and norms.

These dynamics become especially complex when questions about defederating from other servers arise. Many moderators and admins noted that a lot of their defederation decisions are very simple, because it's immediately clear that the remote server is unmoderated or under-moderated by mainstream Fediverse standards—hosting, for instance, hateful and violence-inciting content, extreme gore or pornographic content, spammers, and abuse campaigns. A representative comment from a moderator at Wandering Shop:

> *...if we're going to defederate a server, there's always discussion. And we would defederate a server either for technical bad behavior, like if it's generating spam, or DDoS, or it's a weird somebody's rolled their own Fediverse server that's doing something odd, that might get it defederated.*
>
> *Or if it's a completely unmanaged lawless server, you know, servers that have spun up like Gab and Truth Social and stuff like that was just right on the blocklist. Basically, if it exists to be hostile to other humans, that does it. We deal with the content—my primary focus is, "Is the content problematic?" We will take the minimum action necessary to deal with the content. If the content is problematic because there is an entire server out there that is doing nothing but crypto spamming, we block the server.*

But beyond these easy decisions, questions about when to limit or silence servers require significant moderator and administrator time and attention and reveal underlying but often unexpressed philosophical differences about the best way to think about—and act on—server operators' responsibilities to their members.

2. Whether and when to limit and silence other servers

> *There's still racism and bigotry and homophobia and TERFs and so on out there. But this culture of hair-triggered defederation, much in all as it drives some people crazy, I'm enthusiastic about it. I think it's the right way to go.*
>
> *—Tim Bray, a CoSocial.ca founder*

Despite the fact that most of the server teams we spoke with maintain robust lists of limited and suspended servers, we heard a range of nuanced perspectives on the complexity of decision-making about federation with servers that weren't unambiguously harmful, but which moderated in ways that were in conflict with local norms. Some admins choose to err on the side of maintaining their members' connections to other servers when possible. An admin at Woof.group told us:

> *[Our server] takes a more liberal stance, in that we think that there should be lots of communities with varying community norms and stances, and that's okay. We're not looking to enforce [local] norms on other instances very often. And I think that there's an alternative view of the Fedi, which is looking for much more coherence. It wants a normalized set of content warnings. It wants a certain standard for how interactions go. And if another instance doesn't enforce those norms identically, that instance, anybody who talks to them might be considered bad. We try to ground out in individual conflicts, individual people, instead of worrying too much about policy alignment. And that's not to say we don't have the same goals—we're all interested in anti-racism and building a queer-friendly community, but the way that we go about that is different, I think, [than from] that part of Fedi culture.*

A moderator at Wandering Shop spoke about moderating members of remote servers:

> *There are a lot of cases where somebody says something offensive, possibly even, you know, grievously so, but it's rhetorical, in the heat of the moment, they misspoke. And I don't feel that we should be censoring communication or speech as moderators. I think we need to look at—Is it a problem? Is it causing harm to our instance or to our users? And what is the extent of that harm, or potential harm that we feel we should deal with on behalf of the community?*
>
> *If it's a case of, you know, there's a wound-up scared kid out there, saying something inappropriate about one side or the other in a conflict that's affecting someone, you know, the minimal action would be just okay, take that, take that post down, because that's what's upsetting people. ...it's not it's not a punishment thing. I think a lot of people think it should be, but not in our view here.*
>
> *...and then it escalates from there. I mean, if the user will not back off, or is targeting people, then that's a limit or a suspend on the user. If it's multiple users on an entire instance, even then I try to reach out to the other admin and say, I've had a lot of reports, is something going on? Or are you able to deal with these? And if not, then, you know, that's where we might silence an instance or limit an instance.*

One moderator on a cooperative server noted that despite their personal preference for using defederation sparingly, their team chose, in an example case we discussed, to limit a controversial server:

> *The whole domain limiting…I wouldn't have done this, [but] I think it's a fair call… Limiting limits visibility, but it doesn't actually sever ties, so it's not as disruptive. And then immediately after limiting, [the admin] started this thread [in the server's private discussion forum], which I think is great, and he said, "You know, if you want to block it altogether, suspend—you can just go ahead."*
>
> *[…]*
>
> *This is the kind of thing we need to navigate, right? …because, you know, we don't know…even good actors may turn into bad actors, right? So, to some extent, it's always a general call and saying, like… "Okay, do I trust this actor now? Do I trust that the incentives as an entity corporate or otherwise will remain aligned?" I mean, it's an estimate. You can always say, "Well, if I'm not 100% sure, then no." And the question is how an instance like [ours] can actually serve both kinds of users.*
>
> *This is what brings me to saying suspension is not a good idea, because that's actually telling users that would like to deal with this bridge that they cannot, right? It doesn't maximize preference… So, I personally made a case…which is like, blocking, suspending, at an instance level would be such a huge hammer. This is the hammer we use for Nazis, right? To have that level of hammer, when interacting with instances which are, you know…very heterogeneous, like any human group…it seems problematic to me.*

We also heard repeatedly that server-level defederation is one of the actions that many teams take to a larger group to discuss before acting, given that suspending a server cuts local members' connections to the remote server's members. A representative comment:

> *For bigger stuff, if we're going to block a server, especially if we're going to block a server that has a few followers already, usually we try to message each other. For most of the server blocking things, things still fall to me, just because officially and originally I was the admin. So when it's more like a big decision of blocking servers and stuff like that, it tends to be on my hands, but we still discuss it.*

Those kinds of group discussions and sometimes formal member consultations (even for more top-down teams) were especially apparent in our interviewees' discussions of Threads federation.

3. Threads federation as a governance stress test

Meta's partial adoption of ActivityPub federation for its Threads service came up as an especially complex decision for nearly all the server teams we spoke with, despite the extremely varied decisions the servers in our sample came to. Nearly across the board, server teams brought up entry of Meta's Threads platform into the Fediverse unprompted as an example of a heavier decision process—the potential Bluesky bridge also came up several times in similar contexts, often alongside questions about Threads.

Multiple admins noted that their teams had issued formal statements—often more than one—about their evolving position on Threads federation, several in the form of blog posts. Several teams also conducted formal and informal community conversations about the surrounding issues, and the established cooperative server we looked at held a full deliberative process to arrive at their policy.

A founding administrator of a server associated with a formal non-profit entity related their team's experience with the debate:

> *The question of federating or defederating from threads was one of these issues, where we kept looking at it and we're like—we don't know what they're going to do. Meta has never been a good player in this space. They have never had anything like appropriate moderation of their own community. We can leave it to users to do their own defederating or blocking, but do we want to do that? I don't know. So we had a conversation and the community turned out to be pretty evenly divided between like, "Don't defederate automatically, let us make that decision as we go." And "Heck, no, I don't want Threads anywhere near me, please defederate right away!"*
>
> *And the longer we looked at the responses, the more clear it became that it was the most vulnerable folks within the network who were saying defederate. And we decided ultimately that that was the direction we needed to take, because we didn't feel like it was right for us to tell those folks, you know, tell us if there's a problem and like, let them be the ones who had to experience the brunt of the problem before we took action. That did not seem like a good response for our community.*
>
> *So we wound up preemptively defederating and wrote a post about our thinking and shared it with the community so that everybody could respond. We got a lot of thank-yous out of it. And then I think probably some of the people who are not active on the server anymore, may have looked for a new home where they could connect to Threads. And we acknowledged that we totally understand that that's going to be the response that some of you have, that you're going to want to go to another instance. No hard feelings, we get it—and that, you know, we all have friends who are deep in the Facebook universe and who may end up on Threads, and we would like to entice them to join us instead.*

A founder of a a cooperative server discussed their team's active outreach process when the question of Threads federation arose:

> *We have a variety of techniques for finding out what goes on among our users, so when a vexed question comes up, like, "Should we federate with Meta?" we outreach to the users and do an online meeting or some polls or things like that, and find out what people think. And then the board makes the call. We have a Discourse server that we use to have conversations on. […]*
>
> *I'm actually kind of optimistic about this, about the arrival of Threads. I think this is a good thing. It is absolutely going to be the case that, you know, the libs of TikTok and people like that are going to try to use threads as a platform to harass the Fediverse. And if Threads can't control that, well, sorry, because we tried, but you're defederated… But my vision is that the integration of Threads into the Fediverse, assuming they really do it, and essentially assuming they really do support account migration off of Threads, then the largest net effect of Threads integration is shining a light on the exit door.*

Another admin consulted with their team and made their decision in stages and in response to new information about Threads' approach:

> *…the first statement was, as of now, it's not really a discussion we can have—as of now, there's not really enough info to make a decision. And when we had enough info to make a decision, we made an update to our statement, which was to block all of Threads. And we explained it because the main argument, the main explanation for that decision was that Facebook and Meta had a history of bad moderation. That we cannot trust them at first, but our trust is open to being gained by them by showing a lot of good moderation position, but it hasn't been the case so far. Right. Great. And I think people were mainly okay with the decision, but I saw someone complain.*

4. The potential of federation policies

An interview late in our project with Jaz-Michael King, Executive Director of IFTAS, included an approach to inter-server relations that crystallized our thinking as we synthesized our hundreds of pages of transcribed interviews and noted how much headspace was devoted to working through federation decisions:

> *One of our early findings is that the servers on the larger end of the small side move towards having more documentation around this sort of thing because they hit that pain point and they figured out they need it. And one of the things that comes up over and over again is moderators will... usually have a set of local moderation rules that are written down. It's like your basic, "Here's the rules of engagement on the server." But they don't have the federation rules written down. They have a model bouncing around in their head, but it doesn't get out.*

King noted that as a convening institution (and not a policymaking one), IFTAS emphasizes the benefits of having a formal policy at all, rather than trying to identify an ideal policy and suggest that server operators adopt it:

> *Instead of having a forthright governance model and trying to push that down into people's teams, we're pushing a federation policy template that simply says, "You write down when you do or when you don't federate. And you hold yourself to it. And if you need to change it, you change it." We've been pushing that in response to those questions, instead of saying, "Well, you should or you shouldn't." Why don't you define when you do and when you don't and be public about that?*

Those comments helped us make sense of a gap between what we were *hearing* from team after team, which was that they were spending disproportionate amounts of time and resource on federation decisions, and what we were *seeing* in most public documentation, which was mostly focused on moderation of local members, and on server governance for the teams who'd established more formal or participatory governance processes. Many servers do post public lists of servers from which they defederate—either within Mastodon/Hometown or on a separate site—and many teams have published posts or pages about their reasoning on the Threads federation question, but the criteria behind non-Threads decisions mostly remain blurry.

In hindsight, this gap shouldn't be surprising—it makes sense that most teams have focused first on the kinds of policies that directly affect their members' accounts and behaviors, and on server governance itself, for servers that operate in more transparent, consultative, or participatory ways. It's also true that the fraught nature of many defederation conversations, and the sometimes irreducible interpersonal and coalitional complexities behind them, may have encouraged some teams in our sample to work more from private judgment than from public policy on federation questions, and to have made it difficult for many teams to settle on a comprehensive policy.

Taking those factors into consideration, we think the diverse and robust conversations and viewpoints on Threads federation could serve as a useful jumping-off point for server teams to

even informally document their philosophies of federation and the policies that flow from those philosophies.

In addition to simplifying at least some federation decisions, having those policies—and their implications for members—clearly explained would go a long way toward making it easier for Fediverse members to choose a server that meets their needs. Clear explanations might also serve as a way to introduce new members to the dynamics of inter-server relations and their importance to members' experience of the Fediverse ecosystem.

Section Five: Tooling

Introduction

This chapter is in many ways a side effect of our research. While our interviews did not focus on tools, software tooling came up repeatedly as an area of concern and of potential improvement. This makes intuitive sense: these tools mediate the work of running a Fediverse server.

In this document we try to capture the overall sense of tooling on the Fediverse, both in terms of what admins and moderators are using right now, and also their frustrations and hopes for future tooling. We necessarily focus on Mastodon and Hometown servers since we limited our sample to servers using those software implementations.

About half of the tools-focused conversation fell into what we are calling "moderation." These conversations were focused on the tools and workflows related to the day-to-day activities of moderators on the servers we spoke with. The other half ranged quite widely, so this document is broken into a Moderation tools section with many sub-topics, and then the rest of the document, which covers everything not in that category, such as account migration, federation controls, and so on.

Key observations

- **Very few servers use Mastodon (or Hometown) itself to communicate about moderation.** Usually this happens in an out-of-band channel like Slack, Discord, MS Teams, Signal—our interviewees reported using collaboration software or group chat software of all kinds.
- **All server teams in our sample defederate from problem servers—many very actively and aggressively—but many admins expressed that they don't trust shared blocklists/deny-lists as they currently exist.** Several respondents said they would subscribe to moderation actions from trusted sources via an inbox they could review but weren't interested in fully delegating their blocklist to a service. Some consider blocklists as a useful starting place for a brand new server to take care of obvious bad actors, but not something they need to use in an ongoing way. No admins we spoke to were unreservedly positive about shared blocklists.
- **There is a strong desire for the federation of moderation actions themselves.** Sometimes this was put specifically and technically as a desire for moderation actions to be federated like any other message on the network. Sometimes, as noted above, it was expressed as a wish for a separate inbox where admins could look at moderation actions taken by servers they trust and choose to act on them or not.

- **There is not enough support for formal communication channels between servers.** Inter-server communication is a form of diplomacy and there are currently very few mechanisms for this diplomacy built into the moderation software.
- **Lack of support for account migration complicates moderation decisions.** Since Mastodon provides a limited form of account migration, the promise that if you don't like the policies on one server you can pick up and go to another server rings hollow to users. This reverberates up to moderators who feel like they need to be more delicate with their moderation actions on local users.
- **Financial and legal compliance are areas where third-party assistance could benefit the ecosystem greatly.** There is no financial or legal tooling, either technical or informational, provided by the core software projects. These are areas where many admins feel lost and without guidance.

1. Moderation tools

Discussion of moderation tools dominated our interviews whenever tooling came up. We've attempted to break this large topic into smaller chunks, and there is a lot of crossover and useful dialogue with Section Two: Moderation.

1.1 Documentation and onboarding

A repeated complaint was the lack of built-in onboarding for moderators. While every server will have different moderation policies, and the larger servers we spoke to uniformly had some level of documentation around moderation policy, moderators we spoke to wished that there could be at least some shared documentation for the moderation *tooling*. Some moderators found it easy to learn the basics but it took them a long time to learn the specific nuances of how the moderation tools worked.

There is some movement in this direction by third party entities—IFTAS currently offers a content library which includes both generic and Fediverse-specific moderation documentation.

1.2 Dealing with volume

The Mastodon moderation interface is the user interface where moderators take action on individual reports made to the server, whether from local users or remote users on other servers. It is akin to an inbox containing summaries of each incoming report. Moderators must click on each report, read a summary of the complaint, review the material being complained about, and then pick one of several actions to take.

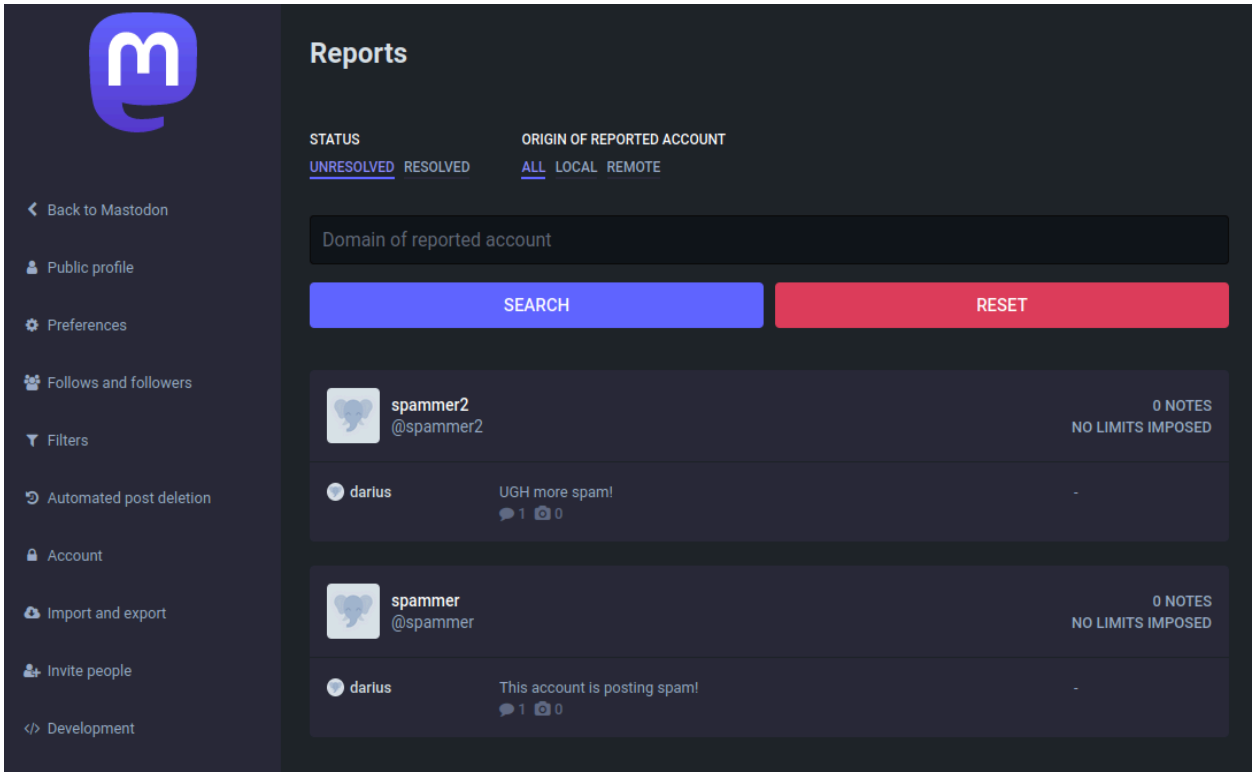
Moderators' ability to manage spam influxes via this interface came up multiple times across multiple sizes of server. A major pain point was the lack of bulk select-and-resolve capability in the Reports page, which is the main moderation inbox. One moderator tells us:

> *The minute someone actually spams the Fediverse, I mean, we don't have tools. We have to use the Mastodon interface to review each report. It took a lot of time. We got like hundreds of reports. If we got thousands, it's like, no, we need scripting. And this is why we started talking to [our technical group], that we want to have a script in place to say, just auto-approve all the pending moderation, for example, like something hacky like this, because Mastodon doesn't allow you to select more than one report at a time.*

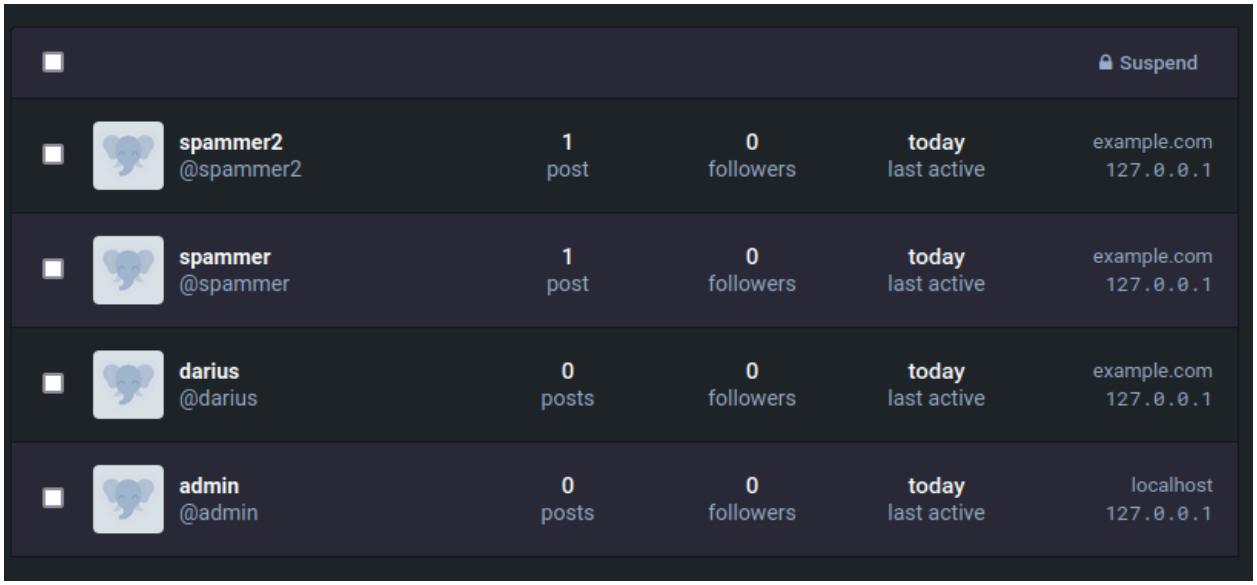
Another admin described the number of clicks required to mitigate spam, saying that for each spam report they would:

> *click, suspend, limit instance, move on. So, yeah. We just determined that it was spam, same as everybody else, and determined that there was no available tooling to really mitigate it. So, just click four buttons a bunch of times. I don't know.*

The same admin described how they figured out a hack where they could use the Accounts interface instead of the Reports interface, sort accounts by creation date, then use the bulk suspend interface there to suspend any recently created accounts that looked like spam. Notably they only applied this to accounts with "nonsense strings" in their names, since there was not a good way to determine from that view whether an account was truly a spammer.



Above is an example of the report moderation inbox in Mastodon v4.2.9, demonstrating the lack of bulk selection.



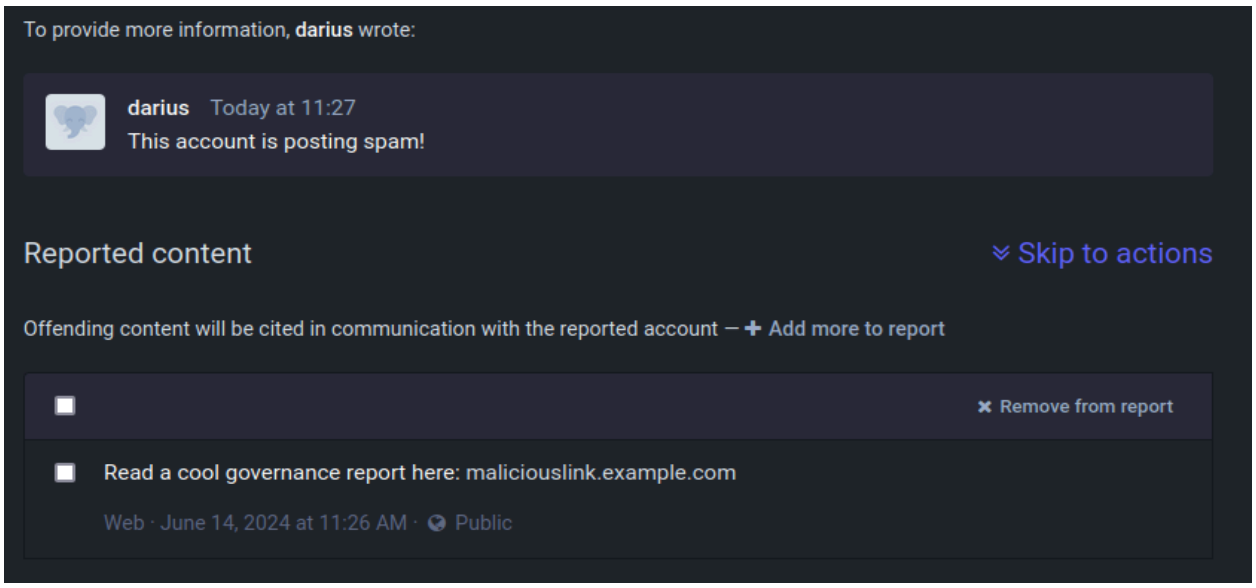
Above is an example of the account moderation interface in Mastodon v4.2.9, demonstrating a bulk selection mechanism and a "Suspend" button in the upper right, which the admin resorted to to manage a spam wave due to the lack of bulk selection in the reports inbox.

On Wandering Shop, which has about 1000 active users, the moderation team consulted with the technical team and used the Mastodon API and other data sources to integrate moderation-relevant events into Discord via webhooks. As a result, they have a Discord channel that aggregates disparate events like reports and emails to the admin inbox in a single place for moderators to view and act on.

1.3 Lack of context

Another problem with the moderation interface that came up multiple times was a lack of visible context for reported content or accounts in the reporting interface itself. When a moderator clicks on a report to decide how to act on it, the only content they see is the content that was reported by the user. In order to see what a piece of content was replying to, or how it was replied to by others, the moderator must click to go to the server of origin and review everything *in situ*. While reviewing comments in their original context can be helpful, it also slows down the pace of moderation. One moderator told us:

> *When a post is reported, it doesn't come in in the context of the thread that it's in, so we have to pull up that thread, assuming it hasn't been deleted or partially deleted by the time we go and check. And the reason it's a problem is, someone might be reported for telling some user to fuck off, and then you pull off who they're telling to fuck off, and maybe they needed to be told to fuck off. You don't know. And so you manually check it. If the post or two above and the post or two below a reported post came through with it, that'd be just amazingly lovely, leaving just that little step would make things so much easier. It also makes it easy to identify if it's just conversation gone wrong [...] versus someone who's being antagonized or harassed. And usually seeing the whole thread makes that clear if it's intact.*



Above is an excerpt of the Mastodon detailed moderation interface in v4.2.9. A moderator can add

citations for further context and offending content, but the context is not provided and the moderator has to click the date-link below the reported content to see more on the server of origin.

1.4 Collaboration between local moderators

Some moderators wished for collaboration features built into the moderation interface.

> *You can't easily track who does what and you can't easily discuss a case with someone else in the moderation team. You can add notes [but] I think the interface towards that is kind of bad. Maybe we could add [...] a reaction like a thumbs up or thumbs down. That could help people make a decision and make a statement as moderators. Because you would know that you would be backed by someone. As of now we have to discuss the case on our Signal group.*

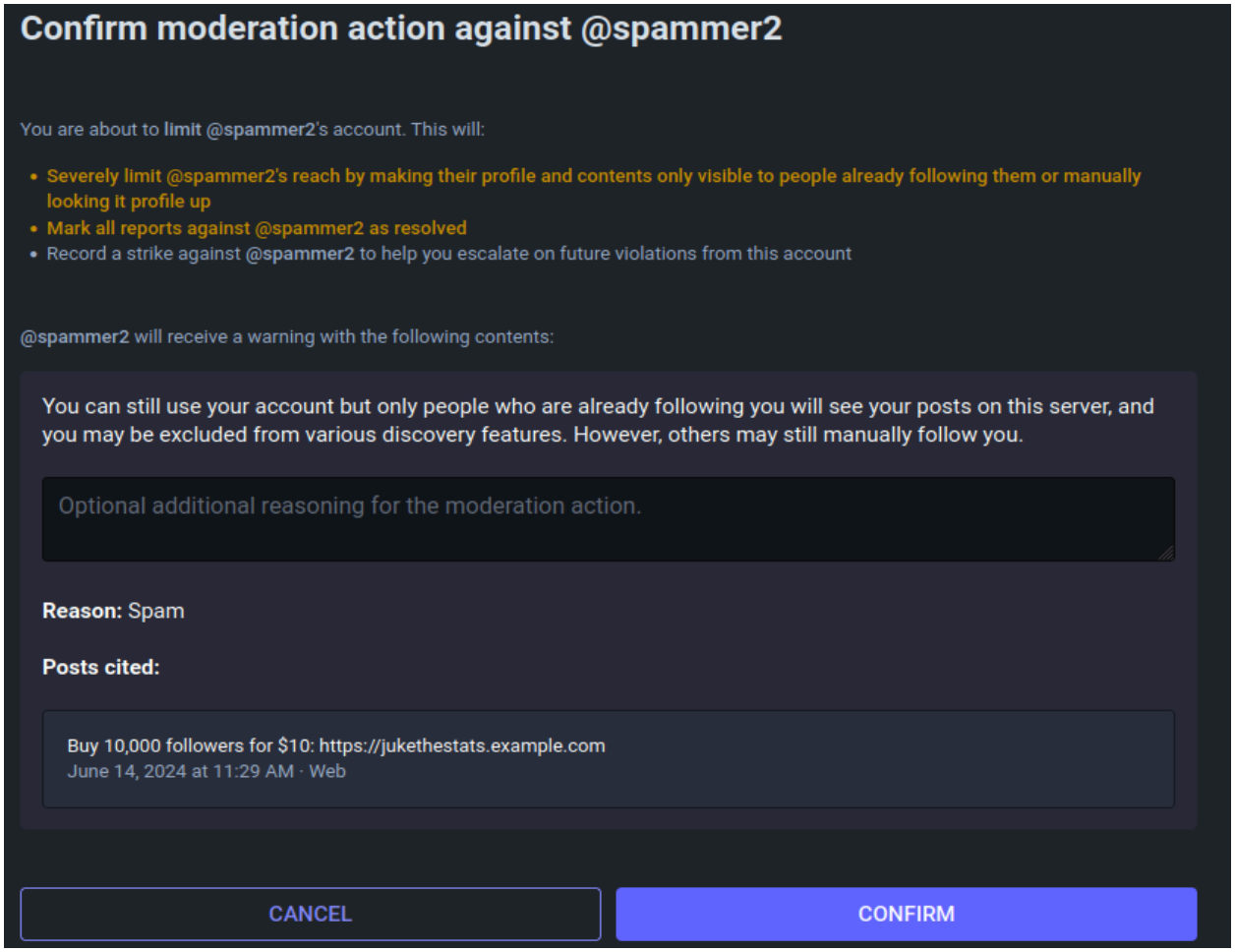
A separate but related issue was moderators being able to collaborate *across servers*—this is addressed later in this section.

1.5 Communication between moderator and user

Moderators wished for ways to communicate with users about the resolution of a report or an appeal. When a moderator resolves a report by taking some sort of action, or by taking no action at all, the user who made the report does not receive any communication that this occurred. According to one moderator:

> *There's no direct way to communicate through the moderation interface itself. Say you want to ask somebody for clarification, or you want to explain why you chose to limit their account, or any of that, there's no real option to do that. You can put a note on when you limit somebody's account, but they can't really say, oh, this is not what I meant, they can't talk to you. And so sometimes people have ended up messaging [a moderator] in a public timeline, and just, this is not where you want to do this.*

While this kind of contact is not necessary for reports filed by remote users who have not agreed to the terms of the server receiving the report, moderators may want to have the option of engaging *local* users to say what happened. At the moment, moderators have to direct message users outside the moderation interface, which adds many extra steps to that workflow and discourages communication.



Above is an excerpt of the moderation interface in Mastodon v4.2.9. There is an option to communicate further with a (local) user who is being moderated, but no option to communicate anything to the original reporter of the issue.

A complicating factor here is moderator safety. According to one moderator,

> *amongst ourselves, we've sort of agreed that if there's a moderation action, we don't tend to go back to the reporting user in a lot of cases, because [...] you don't get a vote in what's done about it. And in general we try to protect the moderators. I may DM someone back to say, thanks for bringing that to our attention, it was a larger problem. And we've done this. But I'm cautious about doing that in a lot of cases, because I don't want to be patting anyone on the back for being a timeline vigilante [...] I don't really want to be encouraging some of the reporting behavior.*

Hachyderm encourages transparency by publishing "Moderator Minutes," a series of monthly blog posts that were intended to be short reads describing in general terms how the moderation team was working and what challenges they faced in the last month.

1.6 Shared blocklists and shared blocks

Many of our participants had strong feelings about shared blocklists (also referred to as deny-lists). Shared bocklists in use today usually take the form of lists of servers, or sometimes lists of users, that meet some threshold for being bad actors in the Fediverse as defined by the people who maintain the blocklist. The blocklists are mostly manually curated, sometimes by individuals, sometimes by groups. The output of a blocklist is usually one or more CSV files that can be imported directly into Mastodon to limit or suspend servers *en masse*. Some blocklists are privately passed between moderators. Others are websites with databases that can export CSV files based on a variety of sub-thresholds that a user can set ("only give me the top 10% of bad offenders on your list," etc.). Some public shared blocklists include the Oliphant.Social Mastodon Blocklists, The Bad Space, and Wesley Aptekar-Cassels' list of large servers with open registration.

The utility of blocklists in the first few months of a server's life is echoed by other participants. According to the admin of one of our core servers with about 300 active users:

> *When we got started, we picked up a blocklist from someone, and I cannot remember who we initially picked it up from. And since then, we do pay attention to the ways that other servers will report instances that they're blocking. We listen to some of that chatter, but we haven't really, I don't think we have picked up anyone's blocklist since that time.*

Several respondents said they would subscribe to moderation actions via an inbox they could review but did not seem open to fully delegating their blocks to a service.

Shared blocks (as opposed to block*lists*) are individual recommendations passed between users and moderators. One form of sharing blocks is #FediBlock, a popular hashtag that individual actors can use to promote servers or accounts they believe should be subject to moderation for one reason or another across multiple servers. The hashtag was created by Fediverse server moderator Marcia X and popularized by longtime Fediverse user Ginger. In a December 2023 interview Marcia X describes its origin "as a tool made by queer femmes to put the spotlight on a sexual harasser."

One admin was supportive of #FediBlock in its early days,

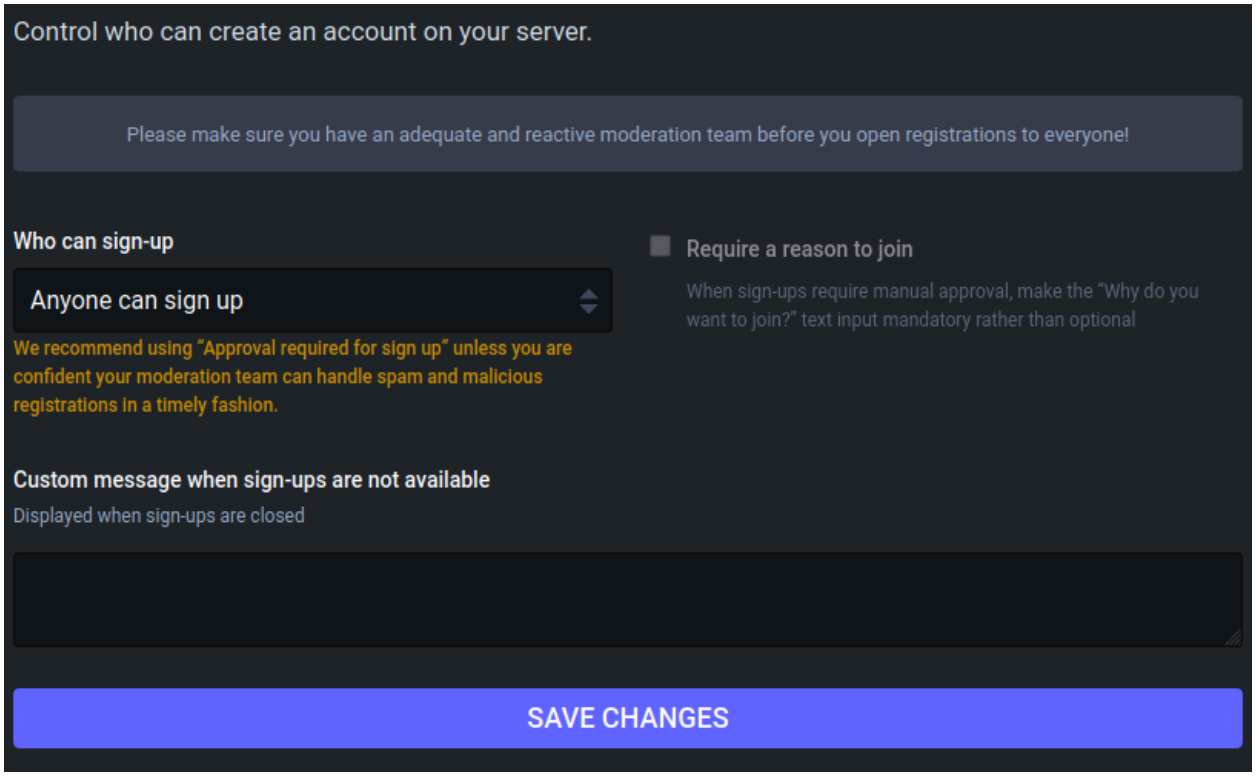
> *but what I found is after years of watching #FediBlock is that the signal-to-noise ratio is extremely low. It is almost impossible [to determine if a recommendation is justified]. I've repeatedly spent like a dozen hours tracing through the partial view of replies because everybody's blocking each other. Everybody's deleting posts, 90% of the discussion is [vague references], there are no links, and I'm trying to get to the ground truth, and finally I find the thing. And it's like, oh, this is a relatively mild, friendly article. It doesn't sound at all like the #FediBlock discussion. And so it was very strange to see that as the defederation criterion.*

> *So we essentially don't respond to random posts calling for proactive blocks unless it's something really obvious like CSAM or hate domains. And those are easy criteria to act on. And even then, we get so little harassment after we blocked the big ones, which happened in the first few months.*

Many admins we spoke to don't fully trust shared blocklists and blocks as they're currently implemented, or see them as a necessary but flawed tool. Given the prominence of blocklists in public discussions on the Fediverse that center on the needs of smaller and less established servers, as well as those with especially frequently targeted members, these lists clearly serve an important purpose, especially during initial setup and for less hands-on server teams than those represented in our sample. We believe that our participants' feedback, including an emphasis on being able to choose in a granular way whether to accept a given block recommendation and the ability to trace the interactions or posts that justify for any individual block, will be welcome in future implementation of shared blocklists within core Fediverse software.

1.7 Account registration control

Some servers find the default tools for account registration limiting. We describe the existing options in detail in our Moderation section, but essentially, registration can be open to all, moderated, or closed/invitation only. We observed some servers that resorted to external survey forms or plain email applications as a way to introduce finer-grained control over who can register an account on their server.



Above is an example of the registration settings page available to server admins in Mastodon v4.2.9. The warning that recommends a server have an adequate moderation team before going to open registrations is new as of February 2024 *and a change we applaud.*

1.8 User-facing generic moderation account

On all but the smallest servers, moderators consistently employ a pattern where they create a generically-named account like `@mods@example.social`. This account becomes an anonymizing front-end for moderation teams which allows them to communicate with users in such a way that communications come directly from the team rather than falling on a particular moderator. This enables moderators to hand over issues to one another in a way that is seamless to the end user.

The anonymizing effect is also an important safety feature. One veteran moderator with experience on Discord and Mastodon noted:

> *I would not ever use my personal account for anything contentious like [moderation], simply because of my experience on Discord, where it's not unusual to have people threaten death.*

One larger server used a pattern where a generic moderators account follows every individual moderator who has access to the account; this provides some transparency as to who is in the group with access while protecting individual members from being associated with any given message sent by the account.

It might make sense for Mastodon or other fediverse software to provide this feature out of the box. For example, instead of having a bunch of moderators share a single password, there could be a group-controlled account that specially-authorized accounts have access to which can be revoked and granted by an administrator.

1.9 Internal moderation team communication

Most server moderators choose some other piece of software aside from Mastodon to coordinate with one another about moderation decisions. The software of choice varies widely, and the initial software of choice tends to default to whatever the initial group of moderators is already most comfortable working in.

One server team uses a free Slack workspace in such a way where the volunteer moderators and operators of the server also have a Slack login. There are channels for tech, customer outreach, and moderation; membership in a Slack channel is equivalent to belonging to a given team, and work is divided ad-hoc.

Another server team uses a Discord server provided by the institution that sponsors them. The moderators work have access to private channels separate from the other projects sponsored by the institution. However they still have access to general channels should there need to be cross-project communication. And another server runs their own Discord where all members have access, but the moderators have access to a private discussion channel.

Our academically-affiliated core server uses MS Teams since their sponsoring institution has a license and they all work in it every day anyway. Another server used Signal groups, again since the moderators already used the software. The moderator we spoke to advised that Signal is okay for four people to coordinate but if they were any bigger they would want to move to workspace collaboration software of some kind. Telegram was used by yet another core server as a back channel between moderators.

Of note is that there was not much open source software used for these purposes, aside from one server that uses NextCould Talk, part of the NextCloud collaboration platform.

1.10 Content filtering

A content filter is any kind of algorithm that ingests messages that arrive at the server and determines whether to block or flag a message based on certain predetermined criteria. Content filtering can be text-based or media-based.

There exist third party content filters used by large platforms, such as Safer by Thorn which scans media posted to a platform and flags if it matches known CSAM (child sexual abuse material). There are similar services to detect spam, violent/extremist content, and other categories that a platform might want to filter. Most of these services require a hefty monetary or infrastructural investment

to implement, and some of these services are reluctant to partner with groups that are not large, known actors due to fear of a reverse engineering attack on their proprietary algorithms.

Content filters are designed from the ground up with large social media platforms in mind. Using these services requires an enterprise-level relationship between the filtering service and the social media platform. As such, these services are nearly impossible for most Fediverse servers to access. In the case of CSAM detection, IFTAS is trying to bridge this gap by becoming the enterprise partner with Thorn and then providing the Safer scanning service to small Fediverse servers via proxy. Even this proxy model is running up against the basic assumptions coded into the content filter software. Jaz-Michael King of IFTAS told us:

> *Everyone I talk to has a product to sell me. And of course, everyone's product assumes that I have all the media and all the text and telemetry. So we paid for Thorn, we paid for implementation help, and we have to keep reminding them, we don't have any of the media.*

The Safer software works on the baseline assumption that the organization they coordinate with is in possession of the media that they are scanning, which is untrue in the case of IFTAS and their proxy service model. Time will tell if these partnerships bear fruit:the work is still actively being developed and has not yet been deployed.

2. Different forms of federation

Mastodon, along with almost all other Fediverse software, is built on a "permissive" model of federation. Any remote server that wants to connect to my local server is free to do so, and I am then free to block or limit that server if they behave poorly. This isn't the only possible model for federation, as several participants brought up. IFTAS Advisor Jon Pincus tells us:

> *[We need] more flexible approaches to federation. Right now, it's pretty much all or nothing. Okay, you've got limit, reject media, you've got a few additional options, but it is still relatively binary. And having something, Emelia [Smith] calls it a firewall approach to federation, where there's much, much finer control over things, that seems really important to get beyond today's default of "accept all federation requests." That's only going to get so far. It's close to breaking down already. But to approve each individual federation request, well, that way lies madness, even for 20,000 servers, let alone if it scales up. So that's a specific area that doesn't exactly fit into moderation tooling. It's kind of infrastructure improvement that can then enable this new class of moderation tool is how I think of it.*

3. Identity and data transfer

> *It's all overlaid with the frustration of, well, the whole promise was if I didn't like things, I could just move. Oh, that's not actually the reality! It's...kind of, sort of the reality. It's complicated. It's complicated, right?*
>
> *—an IFTAS advisor*

As one might expect of a decentralized social network, identity on the Fediverse is fragmented, often by design. This poses problems when communities want to provide multiple services to their users. One of the technical admins we spoke to described setting up a Matrix chat server for their users. They were at first excited to integrate the Mastodon login with Matrix, but it turned out

> *you still had to create a new [account] on the new server. So you can now sign in through the single sign-on with Mastodon, which was a pain to set up, but you still had to create a new [account]. So if there was a way, like an OAuth service for the Fediverse, and then you can have all these different services behind it, and you have just the one identity, I think that'd be great, because [right now] you need to have a new sign-in for our PeerTube, you have to have a new one for PixelFed, which is so annoying. So I think that's one big piece in terms of tooling, if there was a Fediverse OAuth service you can run.*

The admin is describing the fact that when you offer additional federated services for your users (for example, so they can host a blog, or have a video channel, or host photo albums) those users need to create new accounts from scratch on those services.

Relatedly, one of the great promises of the Fediverse is that a user can "vote with their feet" and get up and move to a new server if they don't like the rules and policies on their current server. It is one of the main differentiating factors from centralized social media: this is not a walled garden, and you can move to a new server without losing everything. However, account migration is not as simple as it's made out to be. While there is a mechanism for a Mastodon user to move to another Mastodon server and bring their followers with them, this does not apply to the content of accounts, personal blocklists, or several other categories of information—and this does not apply to people moving from Mastodon to non-Mastodon software.

Of course, there are many social and technical reasons why a user bringing their content with them from one server to another is not currently implemented. For one thing, it would require backdating content or somehow indicating the content's provenance. But probably even harder to solve is the problem of what moderators would do if a user showed up to their front door asking for an account and brought along their history of 100,000 posts. Would the moderators be on the hook for reviewing all of those posts? What sort of liability would the new server have for all this old content that is being imported into their server?

But the friction here is that there is a *user expectation* of bringing their content from one place to another. Either the expectation needs to change, or the technology needs to change.

4. Conceptual location of tools

A question that kept coming up about tools for governance and moderation of servers is whether the tools should live in the core software (like the Mastodon project) or exist as third-party software that interfaces with the core software.

One server moderator expressed frustration that the core projects move too slowly improving their own tools for admins and moderators, and at the same time won't make the (significant) investment of labor to enable outside parties to do the work:

> *Mastodon's moderation tooling out of the box is surprisingly full featured ... [but] building around it is horrific. Let's take Lemmy, Mastodon and Pixelfed. For software that connects so many millions of people, the development teams are extremely disinterested in connecting with people. And the knee jerk [response] is, oh, time, money. But it's also a very strong streak of individualism. And "I got us this far, I can take this the rest of the way." None of these platforms are moving toward plugin architecture, none of these platforms are willing to.*

A tension here is that while third party software would be able to span multiple Fediverse projects and provide a kind of unified view, if the moderation and governance tools are not baked into the core software, many admin teams won't bother integrating external tools, or simply won't know that they exist. More from Jon Pincus:

> *It would be great if more APIs existed to allow third-party tools to be the tide that lifts all the different platforms because Mastodon's dominance is over, but man, I don't see all these other projects necessarily being able to invest a lot in moderation. So these external tools that can work on everything have a lot of value. [...] On the other hand, if it's directly in the [core software], then that's great for the people who just want to get up and go in. [...] But I honestly don't see the [core software] directly investing in moderation heavily themselves. Based on what else they've got on their plate, I just don't see that that is going to happen.*

5. Financial tools

There is infrastructure and tooling needed on the financial side of running a server. One major category is tools that enable the inflow of money from users to operators such as Patreon, Ko-fi, and the Open Collective software platform (which is distinct from the now-defunct Open Collective Foundation, discussed elsewhere in this report).

Expenses are mostly tracked via free tools like Google Spreadsheets. Only one of the 11 servers we spoke with mentioned retaining a professional accountant, though a few had access to volunteers with accounting experience.

5.1 Cooperative decision-making

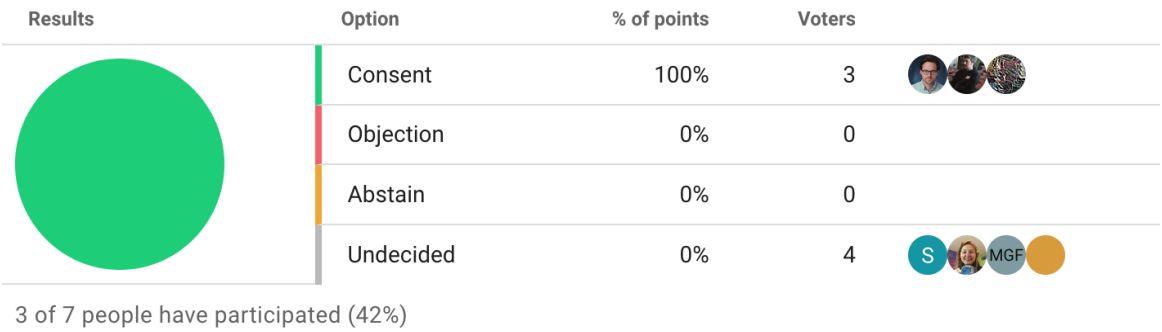
Of special note are some tools for cooperative financial decision-making used by Social.coop.

Social.coop has a Finance Working Group that consists of a handful of members. Within this working group, expenses are proposed, discussed, and approved asynchronously in a dedicated Loomio sub-forum (some but not all forum topics are visible to the non-member public). Loomio is forum software that is designed for consensus-driven decision-making, and is highly configurable with many more forms of voting and discussion coded into its software than your typical forum software. Admins can make very fine-grained decisions about how consensus will be reached. For example a sub-forum can be created where all topics can be discussed for a certain number of days, must be agreed upon via a certain number of people, have certain quorum requirements, etc. Here's an example expense report for Social.coop that has reached consensus via the Finance Working Group Loomio sub-forum:

MGF **Expense 'Mica ops team stipend Jan-May 2022'**

Proposal by Metagov Gateway - Finance · Closed Tue 2022-07-19 17:10

Please vote on this expense submitted on Open Collective: https://opencollective.com/socialcoop/expenses/86458. The vote will end if all group members consent or abstain, or if after six days at least one member has consented. If any members object, the expense will be rejected.



Results	Option	% of points	Voters
	Consent	100%	3
	Objection	0%	0
	Abstain	0%	0
	Undecided	0%	4

3 of 7 people have participated (42%)

You can see there are rules about consent and abstention, as well as a veto rule in place. There is no quorum rule in place. The actual disbursement happens via the Open Collective platform.

5.2 Self-limits on financial capacity

Another interesting financial tooling anecdote comes from Piaille.fr, which has built into their bylaws a limitation on accumulation of funds:

> *I tried to make this as unpowerful as possible.... The non-profit organization, we forced ourselves not to be able to get more than five years of running expenses. So far, we only have banking expenses. So, maybe it's up to 50 euros a year. So, we can't receive any more than 250 euros. Inside our rules, we are due to send back all of the donations once we top that.*

5.3 A gap in fiscal sponsorship

Fiscal sponsorship through the Open Collective Foundation (OCF) was a crucial tool for US-based servers to be able to legally receive funding without resorting to personal bank accounts or incorporation. OCF announced its dissolution in early 2024 with only weeks of notice to the organizations that relied on its services. The guidance given by OCF was for its orgs to find new fiscal sponsors, but at least one server team in our sample that used OCF feels that there are no longer any fiscal sponsors they can turn to. In particular, most fiscal sponsors require some sort of mission alignment, and it has been difficult finding sponsors who consider a general-purpose social media site to be aligned with their nonprofit mission. The server operator we spoke to feels like they have to either return to less structured forms of support or incur major legal costs to incorporate as a nonprofit. Right now they are forced to use a personal bank account while they figure out their next steps:

> *And so today it's all going to a personal bank account.... But it's not ideal, right? We need to move to a more stable structure that's not dependent on any of us personally.*

There seems to be room for at least one Fediverse-focused fiscal sponsor organization in the ecosystem.

6. Legal compliance tools

One small server moderator told us that legal compliance around CSAM is an area where they wish they had both technical tools (for reporting) and public legal guidance tools:

> *I would love to have an automated reporting flow that talks to whatever API the National Center for Missing and Exploited Children does in the US We're legally required to file reports for CSAM. And a big thing I did last year was working with our lawyers to get a letter of opinion about where the lines are, and what we need to report and what we don't. How does caching interact with that? I would love for there to be public, well-vetted legal guidance on what server admins should do—and also integrated reporting. So you can click a post to be like, send this report over with all the metadata required. Because right now I'm filing reports, I'm asking NCMEC for guidance, and they've never responded to me. So I don't know if I'm doing the right thing or not.*

Some organizations like IFTAS are working on CSAM reporting tooling and providing legal guidance around laws like the EU Digital Services Act. But legal compliance in general is a wide ranging area and more tools (both technical and informational) are clearly needed, as admins of small servers feel more or less at sea on these issues.

7. Federation of moderation decisions

> *Having a way to exchange data with other admins has always been a problem on the Fediverse.*
>
> *— a Hachyderm moderator*

As far back as November 2016, the Mastodon project has fielded feature requests from admins asking for some kind of formalized data sharing between administrators. Often this takes the form of requests for formal support for shared blocklists as described in 1.6 Shared blocklists and shared blocks, but other ideas in this space include feeds of moderation actions that servers can subscribe to. According to a moderator of Piaille.fr:

> *The idea [...] was that we could kind of subscribe to a bigger instance or to a moderation instance, which would only publish moderation decisions. [...] You could even figure out some way to subscribe only to a certain field of moderation decisions. [...] But as of now, every Mastodon instance is all alone in moderation. And I think the Mastodon instances could also federate regarding moderation. [...] It would be nice to subscribe to a place that publishes moderation decisions, etc. I was talking about grouping all of the French speakers, administrators, so like we could make big moderation decisions, national or for all the French-speaking instances, etc.*

In part this is in response to duplication of labor: a spam wave is likely to be unwelcome for almost all servers on the fediverse, so why should thousands of individual moderator teams have to spend time investigating and banning the same set of accounts or servers? A moderator of Social.coop explains:

> *#FediBlock is an example. Using a hashtag to communicate blocks, it's—I get it, it's cool, but my question is always, why is this not in ActivityPub? We should be federating moderation actions, right? We should be able to say, like—friend-to-friend instances...we are really friendly, I tell you about everybody we block, you know... I don't think we do it enough, actually... We are being reactive is what I would say. I would love us to do better. Like, when we get a bunch of reports, in particular, of course, when we get forwarded reports from other instances, that's an opportunity to engage with other instances.*

The reactive nature of moderating Mastodon servers came up in other interviews too. A moderator of Woof.group describes what for them is an ideal social structure for inter-server moderation:

> *I want there to be some kind of inter-server moderation discussion channel. Because DMing the other moderators, you never really know if you should DM [certain] accounts, which ones are announcement-only versus monitored, if you go to the personal account [of a moderator] or not. [...] You want to inform a server, "well, no ban happened there, but we do care and what you reported was valid and here's what we're doing about it." And a lot of what I try to do is a sort of political outreach to other instances, to let them know our stance on things. Honestly, it could be just email where we have a norm of monitored accounts and that's where discussions go. It could be something that's built into the server software and it would be cool if there were an inbox that multiple accounts had access to. [...] I would love to be able to see on the report page "Here's the remote instance that reported it" and have a chat system there where you can ask clarifying questions, inform them about the measures you're taking, and so on, and it's all retained in the report log.*
>
> *Because almost all of our inter-instance reports are things that we genuinely care about. We may differ in how we handle them, but they deserve communication. [...] I would love a way to build those sorts of friendly political bridges with other instances in direct relation to reporting.*

They envision a proactive social regime where admins can, in the workflow of moderation itself, run decisions by other known-friendly admins on other servers to get feedback or additional information.

One moderator of hcommons.social speculates that this kind of network-building could result in direct mutual aid across servers, which could essentially share moderation burdens:

> *would allow moderators like me, instead of having to have an entire existence on one server and entire existence on another server, to be able to clock in and help out on different servers or services. That would be fun, because small instances could say, I can't afford a moderator, I don't have anybody, but I want to go away on vacation or something like that they could contact a low cost or free or volunteer, buy me a coffee type service that moderators could help out with. I think that would be nice. I don't think I personally need it right now. But if I decide to go traveling the world for a month, what happens?*

Of course, in order to interact with known-friendly servers, there needs to be a way to flag a given group of servers as friendly relative to your own. A moderator of Social.coop says,

> *we don't have a stream of, like, peer instances.... The question is how to find that beyond just...making friends. Which is nice, clearly. But I don't even know how I would answer the question, "Which instances are in Social.coop's space?"*

Having an inbox of known moderation activities, combined with an understanding of which servers are trusted servers, could let server operators set thresholds for automated moderation ("if X number of trusted servers have blocked this content, I will block it too").

Federation of moderation decisions would reinforce the diplomatic nature of the relationship between servers. Ironically, this type of federation is closer to a classical political definition of "federation" as a kind of resource and information sharing rather than purely publishing things from one person's outbox into another person's inbox.

One stumbling block may be a cultural one. The Fediverse is, broadly speaking, staunchly anti-algorithm when it comes to social feeds. This is at least in part a reaction to the perceived over-reliance on algorithmic feeds on major social media platforms. This may introduce an antagonism toward certain forms of alliance-building tools. Again, the above moderator of Social.coop:

> *Once you start federating [...] decisions and then you move to saying, "Several instances have reported this post, just downrank it, don't show it automatically." That's very, very close to an algorithm. Which is like, I don't know if you've seen, a lot of people don't like the notion of an algorithm because of having been burned by corporate actors. Understandable. But the Fediverse is a bit—I mean, I don't want to generalize, we are like a diverse set of people, but like, a fair chunk of people seem pretty against any kind of like notion of algorithm, which is why Mastodon, I think, lacks some of the tools it needs. [...] We need to have the Fediverse, maybe move beyond a blanket "no" to algorithms, because I don't think I see any other way to scale response. Like what I would love to do is say, yes, we need an algorithm, it means exactly this, and this is what it will unlock.*

This moderator points out that perhaps users should be more open to algorithmic decision making as long as the algorithms are transparent, understandable, and auditable.

8. Inter-server admin comms

Some moderators and admins complained that while they would like to coordinate with (friendly) remote servers to discuss moderation decisions, the pathways and workflows for this are obscure and highly variable. Sometimes moderators form ad-hoc, cross server communities. An admin of Hachyderm tells us:

> *You have to have these little bubbles of contact, and I know in our documentation, we put how to reach us if you're another server admin. Because eventually we were told that server admins were trying to reach us [...] but they didn't use the email address that's affiliated with the server. And so we don't know how they were trying to reach us, but to no fault of theirs. There's not a place in the admin interface for them to go and do that.*

An admin of Woof.group tells us they've experienced similar difficulties with contacting remote admins:

> *I try to do a lot of messaging for anything that's like a non-trivial mod action. I often will DM the moderator on the remote instance. I don't often get responses to that. And I don't know if it's because DMs in Mastodon are easy to miss in the noise, or if it's that we're muted, or they're just busy. I don't know.*

And generally speaking, the results of contacting remote admins were mixed. Most of our participants reported never hearing back when they contact remote admins about moderation or federation decisions. Other participants said they generally hear back a majority of the time. Clearly there is some disconnect here.

The Hachyderm admin went on to discuss a possible solution to the difficulty of admin contact:

> *It would be really nice to see a way for whomever has moderation privileges on the server to have a sort of inbox-y type setup or something between servers, and just let the software, Mastodon or whatever, handle it itself. And if admin communication needs to happen, it can just happen really directly, and you don't have to worry about [forming your own] bubbles. [...] And then you know if your messages are at least being received. And I think that would stir down a lot of inter-instance conflict, which there seems to be a lot of.*

These could also form a basis for communication around other kinds of non-moderation coordination needed between admins. An admin for Woof.group told us that,

> *when another instance goes down or is going to shut down, sometimes we'll coordinate with their mods and make an announcement like, "Hey, if you're looking for a home, we're pretty aligned, just mention where you're coming from and we'll give you an account here."*

This is another place where flagging of known-friendly servers could come in handy, as discussed in the Federation of moderation decisions section. In fact, both the manual inter-admin communications discussed in this section and the automated federation of moderation discussed in the last section would benefit from the same standardized solutions: perhaps a ".well-known" URI provided as an inbox for incoming messages and a "Moderation" activity type that can be used for these communications.

9. Tooling recommendations

In addition to the high-level recommendations near the beginning of this document, we've written an accompanying document, **Fediverse Governance Opportunities for Funders and Developers**, to

collect the many recommendations we've made throughout this findings report. We're also making a series of specific recommendations here that deal exclusively with software/tooling issues in close proximity to the tooling-related observations above.

Recommendations for core software developers:

- **Build pathways for inter-server communication:** In Mastodon, there is no clear path for a moderator on one server to communicate about or appeal a cross-server decision with a moderator on another server. This is a non-trivial feature, which would require attention to safety and consent, but a bare minimum, if each server identified an inbox for inbound admin communications, future collaborative moderation tooling would have a canonical way to send messages intended for a remote moderation team. Standardization of this inbox should be an urgent priority. (Building affordances for both blocklist and allow-list management into this layer or the more ambitious governance dashboard proposed below would allow for more sophisticated and potentially less time-consuming methods of managing these relationships.)
- **Enable allow-list federation:** More rederated projects—including Mastodon—should make it easy for server teams to adopt other forms of federation aside from the "permissive" model now used by Mastodon. Alternatives like allow-list federation are poorly supported, particularly in Mastodon, in which allow-list federation is technically possible but requires setting a Unix environment variable and lacks other first-class UI support.

- **Standardize and enrich tools that control who gets to sign up for an account.** The ability to control account registration was valuable for nearly all the teams we spoke with. Non-Mastodon federated software should consider matching Mastodon's suite of options for registration as detailed above. There's room for more innovation on these features and options within Mastodon as well: some server teams ran more expansive registration processes outside of Mastodon to allow for richer interaction with potential members.

Recommendations for third-party and core software developers:

- **Create a governance-focused dashboard that interfaces with many Fediverse projects:** A governance dashboard distinct from internal-facing content moderation could address both **server leadership** and **federated diplomacy** tasks. This could be a place where mods interact with remote server teams, easing appeals and cross-server discussions, where server leaders can easily create straw polls and other communications to seek input from their members, and where information-sharing alliances such as the coalitional "neighborhoods" or "caracoles" discussed in 3.2 Easing institutions into the Fediverse are formalized via the ActivityPub protocol.

- **Build CSAM-handling tools that are well-suited to Fediverse governance models:** As the Fediverse and other decentralized services grow in popularity, it will be necessary to either

educate third party providers of CSAM-filtering tools (Thorn, PhotoDNA, etc.) and/or build new content filtering software created from the ground up for small servers rather than large companie. Partnerships like the one between IFTAS and Thorn discussed in 1.10 Content filtering are a useful stopgap, but ultimately, if they're going to accommodate new network structures, these providers need to move from an enterprise, handshake-deal business model to a "b2b" model in which the business entity on the other end is a Fediverse server team (or coalition).

- **Develop better moderation tooling:** Many participants described the need for better content moderation tooling that is less manual and labor-intensive, and which includes richer support for communicating with members and other moderators. Developers of core Fediverse software should invest in better tooling to meet these needs, but third-party developers can also contribute to this effort. (One example of more fine-tuned technical infrastructure in non-Mastodon Fediverse software is Pleroma's Message Rewrite Facility, a highly customizable rules system for automated content moderation.) Mastodon, at least, provides API access to their reporting interface, which allows third parties to create moderation tools for or integrate existing moderation tools into Mastodon servers. Building these tools would be a productive area for developers (and funders) interested in fostering better moderation across the Fediverse.
- Core server software should provide the ability for secure, multi-access accounts that are shared by members of their server team who need to communicate with server users. This would allow for a formalization of the creation of generic user-facing administrative accounts like the one discussed in 1.8 User-facing generic moderation account.
- Legal compliance tools, such as automated NCMEC reporting, or settings to handle requirements of privacy laws like GDPR where applicable, would help ease the anxiety of server teams and encourage communities to set up new servers. Third-party projects are a likely place for these tools, as these tools would need to be specialized for the legal needs of different regions of the world.

Section Six: The Case for the Fediverse

The voices of the people who let us ask them hours and hours of questions have been a guiding force throughout this project. In this closing section, we're going to run them without commentary.

I think that as a civilization, the constraints and guidelines and affordances that contain the conversations we have with ourselves is an issue of central importance, and clearly the internet has affected that existentially, the way the humans and human organizations converse with each other. And there have been repeated attempts to guide that through the conduit of privately-owned, centralized, venture-funded capitalist enterprises, and they have consistently failed. Cory Doctorow is right, you know?

I think the jury has been out and it's come in, and we have learned that social online conversations under the auspice of a central capitalist controlling organization is not a recipe for success. And the Fediverse is the best instance I've seen of an alternative that might work, and it's based on the same kind of core structure that email has been. And email is another thing that has managed to survive through all the decades without being monopolized by anybody, and it can't be because you can always get a new email address, right?

So I think the Fediverse is actually existentially important to the quality and success of human discourse—and it's by a wide margin I think the most interesting thing that's going on in the whole human-oriented technology landscape.

—Tim Bray, a founding member at CoSocial.ca

It's my favorite social media! It's interesting, like a lot of people describe this kind of toxicity on Fediverse and I believe them—I think because of the people and communities I've cultivated, and maybe also my identity and things like that, that has not been my experience. When I wake up in the morning and wonder what's on my social media feeds, I'm always much more excited to check what's on Social.coop than what's anywhere else.

...a theory I have is that we're entering a moment where like the VC accelerated social media phase is maybe passing—where all the money is going more toward entertainment platforms. I think of TikTok in that category, it's largely an entertainment platform, and...maybe social media can be boring again. I don't know if this is really true, but maybe AI would just mean like...okay, the AIs are more interesting to doomscroll on than humans.

So let's just make social media slow and really hold it in contrast to that other stuff and not assume it's ever going to make money, you know, and just assume it's this public infrastructure that we use when, when we want to talk to actual people. I fear for the other stuff, but I guess I think of tech as a wildfire—it burns really quickly. And we get a lot of wildfires out here, and there's the front of it, where the blaze is, and then once it's burnt over, that's when cool things start growing up. They grow much slower, and they find their way through the, you know, through the burned trees and new life happens. I kind of hope we're entering that phase of social media that we're done with the fast burn. And maybe it had to happen.

But it's not that old, any of this stuff. And if the rest of the future of social media is not something that VCs think they can make billions and billions of dollars on, fantastic.

—Nathan Schneider, founding member of Social.coop and author of Governable Spaces: Democratic Design for Online Life

My hope is that the technology follows—and if you look at the things like Letterbook, Go To Social, a possible Mastodon fork, these are all things that could move things in a direction that really distills the key value of the sort of queer-centricity, which has largely been lost in the Fediverse as a whole, I would say, because of the huge size. But as this happens, I think there's a chance to restore more of that.

The focus on consent and privacy, I think, is really a distinguishing factor from the directions that Bluesky and all are [taking] and there's a need for something like that. Again, that's not where mastodon.social and all are taking things.... The big wildcard in all that is, can it overcome the whiteness factor? And that's a big question...

Today's Fediverse is a prototype. Prototypes can sometimes evolve into the actual sustainable thing, or sometimes that it evolves into the V1 and V2 happens in parallel, but this is, you know, it's a great test bed for this stuff and both for the instance governance and then even more so for the cross, for the cross instance federation stuff. It's the first time, it's really a prototype at scale that lets us discover these things.

I think the lessons, no matter what happens to today's ActivityPub, Fediverse, you know—the pressures of corporatization are going to lead to huge changes... But the learning is valuable because this core, you know, the queer friendly, indie, we don't want to be commoditized core. That's a network that's going to largely survive in various forms on whatever platform it is.

—Jon Pincus, IFTAS advisor and content moderation researcher

I spend much less time on Fediverse stuff than I did on Twitter. And that has been better for me than not. I explain it to non Fediverse people, it's like methadone versus heroin or something. It feels pretty good, but I don't feel quite as compelled to check it all the time. And that's probably for the best, you know? People are like, "Oh, is it cool? Do you like it like Twitter?" I don't like it like Twitter. And that's why I think it's better...

I think the thing that unites most of our team on this front is we think it's important and a real good, even though the software in this space is pretty clunky, to have people have a non-commercial option for these things. And an option of not being shovel-fed rage bait, basically.

—Phil Siino Haack, advisor to SFBA.social

We've all seen a lot of things rise and fall quickly and slowly. And as someone who's just—I'm so in love with the web and what it means for us as a species, as an animal with eyes and a brain and we get to share and talk!

And I would hate for us to lose this opportunity to reset the last 15 years of walled-garden, surveillance, data-mining weirdness that we milked out of this technology. And return it to the ability to offer folks the freedom to associate and freedom to civil speech…because we're just too cautious as a community to let more people in, to enjoy the thing we built.

…at some point, I think we need to break the glass and understand that the protocol and the platforms exist so that people can associate, connect, communicate in whatever fashion they want to, whether that's a group of two people, 10 people or a million people. There are safe spaces and we need to preserve those safe spaces. But I don't see it as a binary proposition. Not that I'm saying we should make unsafe spaces, but there's a tremendous amount of steam in this engine. And if we don't figure out a way to be proud and expressive about it, I think we'll lose it as well. Because it will just die on the vine…

We're in a very small bubble. We have 15 million accounts, 2 million active, something like that. And so we're highly presented with the troubles of the day in our bubble. I would love to see all of these conversations elevated so that there's a story to tell that is meaningful and can be consumed by the folks who don't know why they would enjoy this experience. Whatever that might be. Whether it's a safe space for furries or a massive connection of people around history or whatever it is.

Hopefully [IFTAS's governance/moderation template work] a one piece of a very large puzzle of adding some robustness and integrity, some structure to all of this so it can grow cautiously, guardedly, mindfully, provide everything, all the benefits that we've eked out of it so far, while providing space for the not us…. This is why I curate that map of the world of Mastodon servers. When I saw a server pop up in Tunisia—"Yes, yes!" Only the people in Tunisia should be in charge of how Tunisians communicate with each other. It's not a San Francisco conversation.

—Jaz-Michael King, Executive Director of IFTAS

I think if we can keep enough people active and involved who are good communicators—it's had incredible reach, the social web in general. It's expanded so many horizons. I mean, there are experiences I've had that I never would have thought of without it. So I'm still enthused about that. It's still happening. I still meet fascinating new people.

And at least in the Fediverse, the signal is still very high compared to the noise, whereas, you know, the collapsing legacy platforms are all noise at this point, doing more damage than good. I want to see things like federated social media set a better standard. I think it can, I don't know that it will. But yeah, that's why I'm still here.

—Johanna B.. moderator for Wandering Shop and CoSocial.ca