

**From the holdings of the Department of Special Collections , Princeton University Library,  
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

---

Master's Theses, 2022-2026

Computer Science, 2023

---

2025

# Enhancing Moderation in Decentralized Social Media: A Comprehensive Approach to Community-Level Blocklist Management and Customization on Mastodon

Zhang, Owen Xingjian

Princeton, NJ : Princeton University

---

<http://arks.princeton.edu/ark:/88435/dsp01zw12z873j>

*U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>*

ENHANCING MODERATION IN DECENTRALIZED  
SOCIAL MEDIA:  
A COMPREHENSIVE APPROACH TO  
COMMUNITY-LEVEL BLOCKLIST  
MANAGEMENT AND CUSTOMIZATION ON  
MASTODON

OWEN XINGJIAN ZHANG

A MASTER'S THESIS

PRESENTED TO THE FACULTY OF PRINCETON UNIVERSITY  
IN CANDIDACY FOR THE DEGREE OF MASTER OF SCIENCE IN ENGINEERING

RECOMMENDED FOR ACCEPTANCE BY  
THE DEPARTMENT OF COMPUTER SCIENCE

ADVISER: ANDRÉS MONROY-HERNÁNDEZ

MAY 2025

© Copyright by Owen Xingjian Zhang, 2025. All rights reserved.

# Abstract

Content moderation is critical for maintaining inclusive and civil online communities by deciding on and filtering posted content. While it has been widely studied in centralized social media, moderation practices on decentralized platforms remain underexplored. Community-level blocklists are a key moderation practice among decentralized social media administrators, specifically those running Mastodon, an open-source microblogging software built on the ActivityPub protocol. Blocklists enable moderator collectives to share details of servers of communities worth blocking. However, little is known about how Mastodon moderators perceive and utilize community-level blocklists, and what additional support they may need. To understand this, we conducted (1) content analysis of five popular community-level blocklists and four related moderation tools, and (2) semi-structured interviews with 12 Mastodon moderators. Our content analysis revealed wide variation in blocklist goals, inclusion criteria, and transparency. Interviews showed moderators balance proactive safety, reactive practices, and caution around false positives. They suggested design improvements like comment receipts, category filters, and collaborative voting. We discuss implications for decentralized moderation, emphasizing trade-offs between openness and safety, complex moderator roles, and design strategies for transparency and collaboration.

# Acknowledgement

I would like to thank Andrés Monroy-Hernández for his invaluable advice and support throughout this project. I am also grateful to Sohyeon Hwang and Yuhan Liu for their outstanding mentorship and encouragement. I thank Manoel Horta Ribeiro for serving as my second reader and offering valuable feedback. Finally, I wish to express my deep appreciation to my family and friends for their unconditional love and support.

# Contents

|                                                                                |           |
|--------------------------------------------------------------------------------|-----------|
| <b>Abstract</b>                                                                | <b>3</b>  |
| <b>1 Introduction</b>                                                          | <b>7</b>  |
| <b>2 Related Work</b>                                                          | <b>10</b> |
| 2.1 Content Moderation in Centralized Social Media . . . . .                   | 10        |
| 2.2 Content Moderation in Decentralized Social Media . . . . .                 | 12        |
| 2.3 Blocklists as a Community Tool . . . . .                                   | 14        |
| <b>3 Findings</b>                                                              | <b>16</b> |
| 3.1 Content Analysis . . . . .                                                 | 16        |
| 3.1.1 Method for Content Analysis . . . . .                                    | 17        |
| 3.2 Interviews . . . . .                                                       | 20        |
| 3.2.1 Method for Interviews . . . . .                                          | 20        |
| 3.2.2 Findings for RQ2 . . . . .                                               | 22        |
| 3.2.3 Demo System . . . . .                                                    | 29        |
| 3.2.4 Findings for RQ3 . . . . .                                               | 30        |
| <b>4 Discussion</b>                                                            | <b>36</b> |
| 4.1 Openness vs. Safety: The Tolerance of False Positives . . . . .            | 36        |
| 4.2 Hero or Villain? Exploring Moderators' Influence on Community Dynamics .   | 38        |
| 4.3 Rethinking Blocklist Infrastructure—Designing for Collaborative and Trans- |           |
| parent Moderation . . . . .                                                    | 40        |
| 4.3.1 Supporting Collaborative Moderation Across Communities . . . . .         | 40        |
| 4.3.2 Enhancing Trustworthy Transparency in Moderation Systems . . . . .       | 42        |

|          |                                                                 |           |
|----------|-----------------------------------------------------------------|-----------|
| <b>5</b> | <b>Limitations</b>                                              | <b>44</b> |
| <b>6</b> | <b>Conclusion</b>                                               | <b>45</b> |
| <b>A</b> | <b>Full Table of Content Analysis</b>                           | <b>53</b> |
| A.1      | Blocklists Analysis (Tables A.1 and A.2) . . . . .              | 53        |
| A.2      | Blocklist-related Tools Analysis (Tables A.3 and A.4) . . . . . | 53        |

# Chapter 1

## Introduction

In *The Open Society and Its Enemies* [Popper et al., 2012], philosopher Karl Popper introduces the “paradox of tolerance,” arguing that unlimited tolerance ultimately undermines itself. In other words, ensuring a free and inclusive society requires limiting the influence of intolerant behaviors. This philosophical insight highlights communities’ ongoing challenge: determining which behaviors to accept and which to reject to maintain inclusive and civil discourse.

Online communities operationalize this principle through content moderation, defined as *the process of shaping information exchange and user activity by deciding upon and filtering posted content according to established rules* [Kiesler et al., 2012, Zeng and Kaye, 2022].

Previous studies have extensively examined content moderation practices at three levels in centralized social media: platform, individual, and community levels. At the platform level, moderation typically involves algorithmic systems and paid or volunteer moderators who enforce site-wide policies and manage user-generated content [Gillespie, 2018, Roberts, 2019, Chandrasekharan et al., 2017]. At the individual level, moderation actions such as reporting, flagging, and personal blocklists help users mitigate harassment and unwanted interactions [Crawford and Gillespie, 2016, Jhaver et al., 2018]. At the community level, collective moderation approaches—such as automated moderation bots and shared community norms—empower groups to collaboratively regulate their own spaces [Jhaver et al., 2019a, Hwang et al., 2024, Seering et al., 2017].

However, these insights predominantly reflect centralized contexts, leaving significant gaps in understanding moderation within decentralized platforms, such as Mastodon Rochko

[2016]. Mastodon is a decentralized social network where users join independently operated servers (e.g., `mastodon.social` and `mstdn.party`). Each server, or “instance,” functions as a self-governed community with its own rules and moderation policies. Moderation decisions are made independently at the instance level, enabling communities to reflect local values while complicating consistency across federated servers [Datta et al., 2010, Rozenshtein, 2023, Hassan et al., 2021b, Anaobi et al., 2023, Hwang et al., 2025]. Within this context, community-level blocklists have emerged as important moderation tools. They allow moderators to collectively protect their communities by blocking problematic instances. Yet, little is known about how Mastodon moderators currently manage these blocklists, the tools they utilize, and the additional support they require. To better understand how community-level blocklists are utilized and perceived, we conducted a mixed-methods study with Mastodon moderators, addressing three research questions:

- **RQ1:** What are the common features of community-level blocklists?
- **RQ2:** How do moderators perceive and utilize community-level blocklists?
- **RQ3:** What design features would enhance the usability and effectiveness of community-level blocklist management tools?

To address RQ1, we performed a content analysis of five widely-used blocklists and four related moderation tools. To address RQ2 and RQ3, we conducted semi-structured interviews with 12 Mastodon moderators, exploring their perceptions, usage of blocklists, and insights regarding moderation tool features.

Based on our content analysis, we found existing blocklists and tools differ substantially in their purposes, inclusion criteria, transparency, and accountability (RQ1). From our interviews, we identified diverse moderation styles adopted by moderators, reflecting careful balances between proactive safety, reactive moderation, and reflective caution regarding potential false positives (RQ2). As a further step, we investigate and evaluate

potential design improvements—including category filters, severity toggles, comment receipts, and collaborative moderation functionalities—that could support flexible, efficient, and informed moderation (RQ3).

In summary, we made the following contributions:

- We systematically analyzed current community-level blocklists and related tools.
- We investigated Mastodon moderators’ perceptions and usage of community-level blocklists, and their perspectives on design features for related tools.

# Chapter 2

## Related Work

### 2.1 Content Moderation in Centralized Social Media

Content moderation is the process of shaping digital interaction and communication by approving, filtering, or removing posted content according to established platform rules or community norms [Kiesler et al., 2012, Zeng and Kaye, 2022]. While moderation is often considered primarily as actions taken by platforms, Jhaver and colleagues [Jhaver et al., 2023] argue that it also encompasses active participation by both individuals and communities. They characterize content moderation practices across three dimensions based on the subject who takes the action: *platform moderation*, *personal moderation*, and *community moderation*.

**Platform moderation** refers to moderation actions taken directly by the platforms themselves, typically employing either algorithmic filters or human moderators. Many platforms rely on automated moderation algorithms to implement site-wide policies and efficiently filter large volumes of user-generated content [Jhaver et al., 2023]. Research critically examines these algorithmic systems, highlighting their effectiveness but also frequent transparency issues, exacerbating accountability and trust concerns [Gillespie, 2018]. Alternatively, platforms deploy paid or volunteer human moderators to handle content that algorithms cannot adequately address [Roberts, 2019, McGillicuddy et al., 2016]. Studies reveal the hidden human labor underlying paid commercial content moderation, illustrating ethical concerns such as exploitation, psychological harm, and

stressful working conditions [Roberts, 2019]. Volunteer moderation has also proven impactful; for example, banning hate-driven communities on Reddit significantly reduces platform-wide hate speech [Chandrasekharan et al., 2017]. Additionally, clear communication of moderation decisions by human moderators enhances user understanding and compliance, thus improving overall platform governance [Jhaver et al., 2019b].

**Personal moderation** refers to individual users’ actions to manage their online interactions and experiences, primarily through mechanisms such as flagging, blocking, and employing personalized moderation tools. Research into user-initiated flagging highlights its dual role as both a method of individual participation in governance and as an influential signal shaping platform-level moderation decisions [Crawford and Gillespie, 2016]. Similarly, blocking tools—where users proactively restrict visibility or interaction from specific accounts—serve as personal moderation measures that significantly impact user experiences. Prior studies examine how the widespread adoption of blocking and blocklists enables users collectively to resist dominant moderation norms and create counterpublics [Geiger, 2016, Jhaver et al., 2018]. Beyond these practices, personalized moderation includes advanced tools such as Squadbox, a crowdsourced moderation tool designed to mitigate individual harassment [Mahar et al., 2018], and customizable word filters, empowering users to proactively curate their online environments [Jhaver et al., 2022].

**Community moderation**, the primary focus of this article, involves content moderation decisions made collectively within user groups or communities, prominently seen on platforms such as Reddit, Discord, and Twitch. Prior research extensively explores how communities develop their moderation norms and deploy tools to enforce them. For example, studies on Reddit’s AutoModerator highlight how volunteer moderators collaboratively configure automated systems tailored specifically to their community standards [Jhaver et al., 2019a]. Similarly, research examining Discord illustrates how community leaders proactively adopt third-party bots as another form of bottom-up

moderation, allowing groups to address complex online harms [Hwang et al., 2024]. On Twitch, prior work demonstrates how moderators actively foster community cohesion by reinforcing desirable behaviors and strategically shaping expectations [Seering et al., 2017]. Broader studies on volunteer moderation further reveal how moderators across various platforms conceptualize their roles through metaphors, emphasizing their active role in shaping community norms and values [Seering et al., 2022]. Across these examples, community moderation emerges as a collective effort, with moderation norms, strategies, and tools reflecting and reinforcing shared community values.

It is important to note that all the moderation practices discussed thus far have been studied exclusively within centralized platforms. As decentralized platforms experience renewed growth and attention, it remains unclear how well existing findings apply to these new contexts. Indeed, emerging evidence indicates that decentralized social media introduces unique and previously unexplored challenges for content moderation—challenges that demand further exploration and understanding.

## 2.2 Content Moderation in Decentralized Social Media

Unlike centralized services, which enforce content moderation through uniform, top-down policies established by a single controlling platform [Gillespie, 2018, Rozenshtein, 2023], decentralized systems are built on open protocols—such as ActivityPub [Group, 2018]—that enable independently governed communities (or “instances”) to interoperate while maintaining autonomy [Masnick, 2021]. This structure allows each community to establish rules that reflect its own values and context [Anaobi et al., 2023, Hassan et al., 2021a,b]. This decentralization shows the potential to enhance participatory governance, improve accessibility, and foster moderation practices more aligned with local contexts. For instance, in Mastodon, individual instances independently set rules tailored to their communities’ needs and values, enabling targeted moderation practices that address

harassment and hate speech [Nicholson et al., 2023b, Liu et al., 2025].

Despite these promising opportunities, decentralized social media also introduces notable moderation challenges. Open federation protocols allow content to flow seamlessly across instances, exacerbating conflicts arising from incompatible community norms [Hwang et al., 2025, Kumar et al., 2018, Zhang et al., 2024]. While differences in community norms are common and expected in online spaces [Chandrasekharan et al., 2018, Fiesler et al., 2018], the interconnected yet independent structure of decentralized platforms makes resolving conflicts between communities particularly challenging. Additionally, decentralized platforms currently lack robust moderation infrastructures and effective tools for transparency and accountability, limiting user trust and understanding across diverse servers [Roth and Lai, 2024b].

To address these challenges, moderators in decentralized social media have implemented various strategies, including manually removing harmful content, issuing user warnings, and blocking problematic instances or individual users to sever harmful information flows [Hwang et al., 2025]. However, these manual interventions often lead to substantial moderator workloads [Anaobi et al., 2023]. Emerging tools and practices, such as curated shared blocklists and automated moderation systems, represent promising avenues for alleviating these burdens. For example, collective blocklists help communities coordinate moderation efforts more efficiently, while automated moderation technologies can reduce moderators’ exposure to disturbing or harmful content, such as child sexual abuse imagery [Spencer-Smith, 2025]. Understanding the effectiveness and implications of these community-driven moderation tools is crucial for realizing the full potential of decentralized social media governance.

## 2.3 Blocklists as a Community Tool

Blocklists have historically been studied as tools for individual-level moderation, primarily on centralized platforms. Jhaver and colleagues [Jhaver et al., 2018] found that user-curated blocklists on platforms such as Twitter are critical for personal protection against harassment but noted significant limitations. Users frequently experience over-blocking and express dissatisfaction due to varying perceptions of what constitutes harassment, underscoring the need for more nuanced and collective moderation tools. Similarly, Geiger [Geiger, 2016] described Twitter’s *blockbots* as community-driven, bottom-up moderation mechanisms enabling marginalized groups to collaboratively shield themselves from targeted harassment.

In decentralized social media, especially Mastodon, blocklists assume an even more prominent and distinct role, particularly at the community level. Recent work emphasizes that Mastodon administrators extensively rely on instance-wide blocklists as key moderation tools, marking a shift from individual to collective protection strategies [Bono et al., 2024]. Despite their importance, little is currently known about how these blocklists are managed, adapted, and integrated into broader moderation frameworks within decentralized environments.

Studying community-level blocklists in decentralized social media provides valuable insights into collective moderation dynamics, especially regarding how communities negotiate shared norms, address inter-instance conflicts, and collaboratively enforce moderation policies. According to prior works [Jhaver et al., 2019a, Geiger, 2016], this collective approach is crucial for non-dominant groups and counterpublics experiencing systemic online abuse, as blocklists offer essential protection beyond individual self-defense tools. Thus, examining common features of community-level blocklists (RQ1), moderators’ perceptions and usage practices (RQ2), and potential enhancements for blocklist management tools (RQ3) will shed light on how decentralized communities effectively

coordinate moderation to tackle persistent challenges.

# Chapter 3

## Findings

### 3.1 Content Analysis

On the Fediverse, blocklists are curated lists of domains (instances) that moderators choose to restrict or block from interacting with their own instance. These blocklists are typically created and maintained by individual administrators, moderation teams, or trusted community groups seeking to protect their communities from harmful content such as spam and harassment. They are often shared through public repositories, admin forums, or the creators' own websites. To streamline the adoption and management of these blocklists, moderators increasingly rely on dedicated tools that help them discover, evaluate, and apply blocklists in ways that align with their local moderation values.

To answer RQ1, we conducted content analysis on a set of blocklists and a set of blocklist-related tools widely used by Mastodon moderators to understand the current features they have.

### 3.1.1 Method for Content Analysis

#### Criteria for Selection

Drawing on community-curated lists<sup>1,2,3,4</sup>, we selected five blocklists (Seirdy Tier-0, FediNuke, Garden Fence, CARIAD, and IFTAS-DNI) and four blocklist-related tools (FediCheck, FediBlockHole, Fediseer, and The Bad Space). Our selection criteria included public accessibility, community-driven maintenance, extensive documentation, and widespread adoption within the Mastodon community. We adopted this approach to ensure our analysis reflects real-world moderation practices and captures tools and decisions widely recognized and implemented by Mastodon moderators.

#### Content Analysis Procedure:

For both the set of blocklists and the set of blocklist-related tools we examined, we engaged in an iterative review of each item in the set. To analyze and characterize current blocklists, the first author first thoroughly read the page and documentation of every blocklist in the set. They then created a rough list of dimensions of key information about blocklists that would serve as categories to code each blocklist along. Getting regular feedback from other authors, the first author coded each blocklist along the categories, which led to minor adjustments of the categories to ensure conceptual consistency and alignments across the blocklists. This process was repeated separately for blocklist-related tools, resulting in a different set of categories along which those tools were analyzed. The final categories and their definitions are shown in Table 3.1. The results are shown in Table 3.2 and Table 3.3. Full table in Appendix A.

---

<sup>1</sup><https://codeberg.org/nev/awesome-fediadmin/#basic-blocklists>

<sup>2</sup><https://connect.iftas.org/library/iftas-documentation/cariad-policy/>

<sup>3</sup><https://github.com/ineffyble/mastodon-block-tools>

<sup>4</sup><https://writer.oliphant.social/oliphant/the-oliphant-social-blocklist>

Table 3.1: Combined analytical categories for Blocklists and Blocklist-related Tools

| Type      | Category           | Definition                                                                                                |
|-----------|--------------------|-----------------------------------------------------------------------------------------------------------|
| Blocklist | Purpose            | Initial motivation and ultimate goal behind creating the blocklist.                                       |
|           | Criteria           | Specific requirements determining instance inclusion or exclusion.                                        |
|           | Transparency       | Explanation of review processes, resources used, and how these resources informed blocklisting decisions. |
|           | Distribution       | Public availability and frequency of updates.                                                             |
|           | Limitations        | Limitations or challenges explicitly reported by the blocklist creators.                                  |
|           | Other Resources    | Additional recommended blocklists or tools for moderators beyond this blocklist.                          |
|           | Additional Content | Supplementary contributions or content provided by blocklist maintainers beneficial for moderators.       |
| Tool      | Summary            | Brief overview describing the main functionality of the tool.                                             |
|           | Purpose            | Motivation and objectives behind the creation of the tool.                                                |
|           | Technical Access   | Means through which users interact with or technically access the tool.                                   |
|           | Decision Type      | Methodology or basis by which the tool makes or facilitates moderation decisions.                         |
|           | Tags/Categories    | System for tagging or categorizing moderation entries provided by the tool.                               |

Table 3.2: Summary of Mastodon Blocklists

| Category           | Garden Fence                                                                | CARIAD                                                                            | IFTAS-DNI                                                      | Seirdy Tier-0                                                                   | FediNuke                                                                     |
|--------------------|-----------------------------------------------------------------------------|-----------------------------------------------------------------------------------|----------------------------------------------------------------|---------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| Purpose            | Starter blocklist protecting users from well-known harmful sources.         | Initial moderation resource for widely-blocked domains.                           | Curated subset of CARIAD domains recommended for defederation. | Customizable starter blocklist aligned with Mastodon Covenant standards.        | Recommended default blocklist focusing on severe and notorious instances.    |
| Criteria           | Includes domains blocked by sunny.garden and majority of reference servers. | Combines data from the IFTAS DNI list and major Mastodon providers.               | Labels domains according to DTSP Trust & Safety Glossary.      | Consensus-based inclusion (appearing on $\geq 15$ of 27 source blocklists).     | Severity-based subset, prioritizing severe, notorious actors.                |
| Transparency       | Domains reviewed manually, reference servers not publicly disclosed.        | Includes domains with 51% consensus, manually reviewed by IFTAS.                  | N/A                                                            | Curated by accountable individual, sources selected privately with overrides.   | Manually curated, severity-focused, supported by strict receipt criteria.    |
| Distribution       | Publicly available, updated weekly.                                         | Available via FediCheck service.                                                  | Publicly accessible on IFTAS Connect.                          | Public CSV form, periodically updated.                                          | Public CSV form, updated less frequently and selectively.                    |
| Limitations        | Subjective judgments, English-language bias.                                | Biased toward global north perspectives; lacks marginalized community protection. | N/A                                                            | Subjective, limited transparency, slow manual updates.                          | Inherits Tier-0 limitations, intentionally limited completeness and updates. |
| Other Resources    | Recommends additional blocklists and moderation tools.                      | Suggests tools for synchronizing Mastodon blocks (e.g., FediBlockHole).           | N/A                                                            | Suggests FediBlockHole, supplementary blocklists.                               | Same as Tier-0.                                                              |
| Additional Content | N/A                                                                         | Public feedback channels provided.                                                | N/A                                                            | Documents moderation evidence ("receipts"), manual updates, community feedback. | Same as Tier-0.                                                              |

Table 3.3: Summary of Mastodon Blocklist-related Tools

| Category         | FediCheck                                                                       | Fediseer                                                                | FediBlockHole                                                                  | The Bad Space                                                                                   |
|------------------|---------------------------------------------------------------------------------|-------------------------------------------------------------------------|--------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| Summary          | Moderation-as-a-service, automatically syncing blocklists for Mastodon admins.  | Crowdsourced human-curated spam/harassment classification tool.         | Open-source tool auto-syncing Mastodon server blocklists from trusted sources. | Community-driven catalog identifying harmful instances via marginalized communities’ consensus. |
| Purpose          | Helps new admins quickly implement moderation by auto-applying vetted denylist. | Human-curated ”chain of trust” for spam classification.                 | Simplifies moderation by aggregating trusted blocklists.                       | Lists servers posing risks due to poor moderation or harmful content.                           |
| Technical Access | Web app (Mastodon v4.1+ API) for remote domain block management.                | Machine-readable API to access spam classification.                     | Service/script via Mastodon Admin API for fetching/pushing blocklists.         | Web-based searchable directory, public API for flagged instance data.                           |
| Decision Type    | Automates blocks using aggregated and moderated lists (CARIAD/DNI).             | Uses three judgment types: guarantees, endorsements, censures.          | Allows subscriptions to external lists with admin final control.               | Lists servers only after independent decisions by multiple communities.                         |
| Tags/ Categories | Uses standardized DTSP labels for classifying behaviors and content.            | Allows up to 100 descriptive tags per instance for filtering/discovery. | Does not provide own tagging; supports public/private comments.                | Categorizes reasons by harmful behavior types identified by communities.                        |

## Findings for RQ1

Our content analysis reveals that current Mastodon blocklists serve primarily as foundational moderation tools but exhibit significant limitations in transparency and consistency. Related tools offer technical facilitation but often lack standardized mechanisms for community feedback and content categorization.

**Blocklists.** The blocklists analyzed typically function as a ”starting point” moderation resources, offering new moderators a foundational set of blocked instances to mitigate initial moderation burdens. These lists universally emphasize cautious, manual review processes, highlighting moderators’ strong preference for careful, human oversight in deciding blocklist inclusion. However, these blocklists vary considerably in their focus, ranging from consensus-driven to severity-based decisions. They also carry inherent limitations: decision-making processes are often subjective, biased toward particular values or linguistic groups (e.g., English-speaking communities), and lack consistent transparency, leading to potential concerns about trustworthiness and timeliness of updates.

**Blocklist-related Tools.** Blocklist-related moderation tools primarily support moderators through technical mechanisms such as web interfaces, scripts, and APIs,

significantly easing administrative workload. These tools assist moderators throughout different stages of moderation, from making initial blocking decisions and importing existing blocklists to subscribing to external sources and updating instance-level moderation rules. Despite their utility, these tools commonly miss essential features such as consistent tagging systems and integrated feedback mechanisms, limiting moderators' ability to revise or contextualize moderation decisions collaboratively and adaptively. These insights underscore critical gaps and opportunities in current moderation practices on Mastodon, pointing toward potential design improvements for future moderation tools that emphasize transparency and community collaboration.

## **3.2 Interviews**

To answer RQ2 and RQ3, we conducted semi-structured interviews with 12 Mastodon moderators, focusing on how they perceive and utilize blocklists and their perspectives on blocklist-related tools.

### **3.2.1 Method for Interviews**

#### **Participant Recruitment**

To recruit participants, we sent direct invitations via Mastodon direct messages (DMs) and emails to Mastodon instance owners identified through filtering on instance.social for instances with more than 10 users. Additionally, we employed snowball sampling by inviting recommendations from initial respondents to expand our participant pool. Participants who expressed interest in the preliminary demographic survey were then contacted to schedule interviews.

## Study Protocol

The study began with a pre-interview demographic survey and informed consent procedures. Semi-structured interviews were conducted with 12 Mastodon moderators, aged between 26 and 55, from seven countries (USA, France, New Zealand, Netherlands, Canada, China, and South Korea). Participants included 6 males, 2 females, 2 non-binary individuals, and 2 who preferred not to specify their gender. Details are shown in Table 3.4. Each interview lasted approximately one hour and was structured into two segments. The first 30 minutes explored participants’ current practices in using and managing blocklists (RQ2). The second 30 minutes introduced our prototype moderation tool and focused on moderators’ specific needs and preferences regarding moderation tools and features (RQ3). Interviews were conducted remotely via Zoom, audio-recorded with participant consent, and participants were compensated with a \$15 payment via Zelle for their time. This study was approved by the Institutional Ethics Review Board at our university.

Table 3.4: Participant Demographics

| ID   | Role          | Country     | Gender            | Instance Registration            | Active Users        |
|------|---------------|-------------|-------------------|----------------------------------|---------------------|
| P1   | Owner         | France      | Male              | Open (by admin approval)         | 300+                |
| P2   | Owner         | USA         | Female            | Closed                           | 100+                |
| P3*  | Owner         | New Zealand | Male              | Open, Open (by admin approval)   | < 100 (3 instances) |
| P4   | Moderator     | USA         | Prefer not to say | Open (by admin approval)         | 300+                |
| P5   | Administrator | USA         | Non-binary        | Open (by admin approval)         | 700+                |
| P6*  | Owner         | Netherlands | Male              | Open                             | 200+, 9700+         |
| P7   | Owner         | Canada      | None              | Open (by admin approval)         | 500+                |
| P8   | Owner         | USA         | Male              | Open                             | 100+                |
| P9   | User          | USA         | Non-binary        | Open (by admin approval)         | 100+                |
| P10  | Owner         | China       | Prefer not to say | Open (by admin approval)         | 600+                |
| P11* | Owner         | Canada      | Female            | Closed, Open (by admin approval) | 100+, 800+          |
| P12  | Owner         | South Korea | Non-binary        | Open                             | 100+                |

*Note:* Participants marked with (\*) moderated multiple instances. Moderator” includes Owner, Administrator, and Moderator roles, as all have moderation access. P9 was included as an experienced Fediverse writer and is an exception.

## Data Collection and Analysis

Data were collected from audio recordings of the interviews and survey responses. All interview recordings were transcribed verbatim and stored securely, with personally

identifiable information redacted to ensure participant confidentiality. The first author independently conducted thematic analysis by developing a codebook and coding all transcripts, enabling detailed qualitative insights relevant to our research questions and exploratory goals.

### 3.2.2 Findings for RQ2

We consider blocklists as the combination of the database of instance information **and** the moderation decisions people make to include or exclude each instance. Hence, to answer RQ2 (How do moderators perceive and utilize community-level blocklists?), we broke it into three questions: what are people’s moderating styles, how do people make blocking decisions, and how do people think and use blocklists.

#### What are people’ moderating styles?

Participants demonstrated diverse moderation approaches, categorized into five styles: (1) report-and-fix, (2) find-and-act, (3) exposure-first, (4) safety-first, and (5) careful-decision.

**Report-and-fix.** Moderators adopting this style rely heavily on user-generated reports to trigger moderation actions. P4, a moderator, described the daily moderation task as “either for a join request or an account report”—moderation happens when there is a report. Although it is a reactive moderating style, P11, who manages 2 instances, encouraged users to report, “if you’re not sure if you should report it—report it. And I’m just going to tell you, we will do something, or we won’t, based on the content of the report. ... we have a line in our code of conduct that says, ‘Giving the admins a headache can get you moderated.’” P8 also valued the user input—it is possible to revert their blocking decisions if their users ask about it, “because at the end of the day, we only have two or three hundred active users. We don’t have a ton of people, and we really value the direction they want to go.”

**Find-and-act.** This moderation style involves proactively identifying and blocking

problematic instances to reduce future workload. P10 mentioned that they block instances that do not regulate right-wing content and hate-speech, but doesn't have a "systematic way to detect and react to such servers", so it is more like a "find-and-act style" to manage blocklists. P4 also thinks there was a desire to be more proactive in content moderation, especially in "an early stage of becoming a moderator and as the community started to grow". P5 noted that this find-and-act is an action that save time for future efforts, "because if I didn't block these instances, then a few months later, their users would harass our users, and then we'd get a lot of reports, and I'd have to block them anyway. So I think a proactive approach is actually good for a lower workload", P5 said.

**Exposure-first.** Moderators prioritizing openness emphasize exposure, and they are worried that over-blocking would hurt innocent people from underrepresented communities. As P8 described, "I think exposure is really important. Just like in New York, where you walk outside and you're exposed to everything. ... I try to embody that culture a little bit in the server." And P4 expressed worries about over-moderation, thinking it hurts marginalized communities.

"I'm very sensitive to over-moderation. ... One big issue is that people will brigade accounts that are doing nothing wrong, simply because they don't like them. This is how many queer people and others from marginalized communities end up getting banned on social media—right-wing users mass report the account, and moderation systems typically ban it first. ... I'd just worry about something similar happening here." – P4, LGBTQ+

**Safety-first.** Participants argue that over-blocking is safer than under-blocking, since the risks to your own users are far lower, and it would prevent some mental burdens for new moderators. P9 said,

"I have to say, my general philosophy is that I'd rather over-block than under-block. If I block an instance by mistake, it's not the end of the

world—people have various ways to get in touch with me. They can reach out and let me know, or if someone on one of my instances asks, 'Hey, why are you blocking them? I've got some friends there,' then I can go look into it. ... At most, there's some temporary loss of communication. But to me, that's far less risky than not blocking something that could lead to harassment." – P9, wrote 100+ articles about Fediverse

Although P8 wants more exposure for their instance, they suggest new moderators to be more restrictive,

"I would say to a new moderator: feel free to start too restrictive and work your way back to your comfort level, because burnout is real. Just find the sweet spot for what you can manage. Moderating can feel like boiling the ocean—it doesn't have to be that way. ... If you want to, you could open up your server to everything in the world. You could promise every safety measure and choose not to block any servers. But if you go that route, it's going to be a lot of effort. The most important resource in a Trust and Safety role is yourself—your own energy, your mental health." – P8, 3 years experience managing an instance

**Careful-decision.** Participants demonstrated a high level of awareness in making moderation decisions based on varied reasons. Both parties, those favoring exposure-first and those favoring safety-first, think careful decisions are very important. As one who advocates for exposure, P8 mentioned that blocking a whole server is like the "last resort," and as a moderator, these decisions could possibly make your users' world a little bit smaller. P9, who has written 100+ articles about the Fediverse and prefers over-blocking to under-blocking, suggested doing manual reviews, saying, "I think the best practices for an instance block list are to manually review everything. (Though) that's a lot of time." Other than that, P3 and P12 also stated that instance blocking is a "huge impacting decision", where moderators should take "serious responsibility" since it affects whole instance's connectivity.

## How do people make blocking decisions?

Based on our interview results, we decomposed the decision-making process into four parts: (1) how moderators locate instances to block, (2) what indicators guide their decision, (3) what action they take, and (4) which instances they want to block.

**How moderators locate instances to block.** Two common ways to locate potential servers to block are by following the #FediBlock hashtag and checking user reports. Participants (P1, P5, P6, P9) mentioned they are following #FediBlock hashtag, where many moderators would post the worth-blocking servers with the hashtag #Fediblock, to check if they need to make any actions. Besides that, moderators would notice patterns if they receive multiple user reports about accounts from the same instance, prompting them to check that instance. As P1 mentioned,

“I have two reports from generated accounts by some bots on the same server.

So it’s a kind of clue that, oh, maybe this server is hosting a lot of spam bots and stuff like that. So I need to investigate by myself and check the server.” -

P1, owner of instance with 100+ users

**What indicators guide their decision.** After locating these instances, moderators make their decisions based on indicators such as the homepage, rules, and how the instance’s moderators act. P1 noted that sometimes it is very direct and quick to determine if a instance is problematic by checking its description, admin information, and public feed to get the “vibe,” making it easy to suspend the whole instance. P10 focused on the rules of the instance, stating, “if the rules of the instance are specially catered to a group of right-wing fascists, or do not regulate hate speech at all, the instance, in my opinion, is a block candidate before its posts proliferate.” Building on that, P11 described a series of actions taken to determine if an instance should be blocked: they first suspend individual users, and if they observe that the instance’s moderators are not actively moderating, they then block the entire instance.

“If the instance—our take on the instances is they should be good Fediverse citizens, which means they need to have an active administrator and/or moderator. If we see those things(bad behaviors) happening, then we will suspend a user who’s a problem for us, but the instance is fine. If we get no management from the instance—or the instance admin themselves is the problem—those ones are the ones that we block.” - P11, manages multiple instances

**What action they take.** Participants would block spam instances but would limit the instance or block individual users instead of blocking the instance if it is borderline.

Participants (P3, P5, P6, P9, P11, P12) said it is easy to suspend instances that are full of spam accounts, since these instances are very obvious and they would suspend them very quickly without second thoughts. Other than that, they might use “limit” rather than “block” for borderline instances or block individual users instead of blocking the whole instance. According to P7, there are many cases where a user’s behaviour is borderline—serious enough to suspend the user, but not serious enough to suspend the whole server. P7 also mentioned, “I occasionally use Limit instead of Suspend for borderline cases, where a user is perhaps overly argumentative or peddles in uninformed opinion, but doesn’t step clearly outside the rules of the server.”

**Which instances they want to block.** P1, P5, and P6 “wished” they could block large instances like `mastodon.social` and Threads (`threads.net`). Although many spam and troll accounts exist on `mastodon.social`, blocking it would mean losing the opportunity to interact with a large number of users. As a result, P1 and P5 expressed concerns about these large instances but ultimately chose not to block them. P6 observed that many communities are blocking Threads because they perceive Meta as harmful; however, P6 has not blocked Threads, explaining that “I don’t see any reports coming.” Beyond large instances, other participants described clear goals for blocking instances that allow problematic content, including CSAM, spam, racism, harassment, and anti-trans behavior.

As P7 explained:

“I typically use server blocks for servers that are clearly outside the bounds of what I consider acceptable behaviour, such as allowing harassment or racism (often under the guise of ‘free speech’), or allowing content that is potentially illegal or containing CSAM.” — P7, author of a popular blocklist

### **How do people think and use blocklists?**

We found participants have diverse opinions on using blocklists. We categorized them into two groups: (1) Opinions on liking or using external blocklists, (2) Opinions on disliking or not using external blocklists.

**Opinions on liking or using external blocklists.** Participants who liked blocklists thought they provides information to users when choosing which instance to join and help block potential threats in advance. P1 thinks that blocklists are a good practice for users who want to join instances because they can explore the list and see why some instances are blocked, especially if there are public comments. Since applying a blocklist is a preemptive action, P4 also mentioned that it is unknown what moderation would be like without blocklists. P10 considers using community-level blocklists as a positive experience for blocking potential intolerant harm and harassment.

Participants also used blocklists and related tools to save time and help create their own blocklists. Among the five blocklists and four tools mentioned in the content analysis, P1, P2, P6, and P11 mentioned their use of the blocklist management tool Fedicheck, and P9 used Seirdy’s blocklist and The Bad Space as references. P1 mentioned that Fedicheck makes moderation much easier by saving time on updating blocklists, “and honestly”, P1 said, “for a small server like ours, I don’t have this time.” P2 also said it does not seem necessary to use other moderation tools and that Fedicheck is sufficient at the instance level. P9 told us they used a couple of blocklist sources when getting started, including Seirdy’s blocklists, and also used The Bad Space, thinking of it as “a kind of instance

catalog.”

Speaking of how they manage blocklists, P1 and P11 mentioned adding private comments and public comments to the blocklist. They pointed out that private comments, or internal notes, are only visible to moderators, helping to explain to others—and to their future selves—the detailed reason behind each decision, making it “justifiable, renewable, and reviewable.” Meanwhile, public comments are more like a technical reason why the decision was made, justified against the server rules.

**Opinions on disliking or not using external blocklists.** Participants who dislike blocklists think managing blocklist is both time-consuming and mentally-consuming, and they don’t trust others’ blocklists. P7 described, “The main thing I dislike about blocklists is the state of tooling for managing fediverse and Mastodon blocklists in particular. There is a considerable amount of manual effort involved in dealing with blocklists.” This manual work makes participants both time-consuming (P1, P7) and mentally-consuming (P12). Additionally, P3 and P9, both who have more than 7 years of experience in fediverse, doubted if people should trust other blocklists. Firstly, it is unknown if the how people curated their blocklists - is it independently verified or they are just taking someone else’s word for it. (P9) Also, for people who independently verified the blocklists, there would be many biases - systematic bias where the referenced blocklist has the different judgement that does not align well with yours, and personal bias like “blocking toward people who speak a different language, or use any number of criteria that have very little to do with the actual purpose of the blocking”. (P3)

For participants who do not use external blocklists, P3 stated that they chose not to use external blocklists because an instance’s behavior can initially be unstable if administrators have not established strong processes, and because instances can change quickly as moderators gain experience and improve their moderation systems. P12 noted they moderate manually but would consider using blocklists if “some neo-nazi instances are popping up like crazy, when I can no longer moderate the servers hand-to-hand, by my own

effort”. Also, for marginalized community, popular shared blocklist are not stand on their sides, P9 explained,

“IFTAS put together their CARIAD blocklist—which is now changing and no longer being distributed as a blocklist—it was originally based on aggregating blocks from the most-used servers. They specifically stated that this blocklist will not protect Black, Asian, Eastern, and trans communities, because that’s not the philosophy of the top servers, which are generally run by cis, white guys.” - P9

### 3.2.3 Demo System

To support our exploration of customizable blocklist management practices, we developed a lightweight web-based demo system. Built with React and Vite and styled using plain CSS, the system simulates key moderation functionalities tailored for Mastodon moderators. It highlights three main features: (1) category filters, (2) severity toggles, and (3) comments. The user interface is shown in Figure 3.1.

**Category filters** enable moderators to selectively view instances based on specific comments, such as spam, harassment, or misinformation. This is displayed on the left-hand side panel.

**Severity toggles** allow moderators to assign different levels of action to each instance—Suspend, Limit, or None—each accompanied by an explanation. These actions are visually represented with red, yellow, and black indicators.

**Comments** provide a space for moderators to check the reason for blocking each instance. When users click on a comment tag (such as the cursor shown in the image), the system displays its definition and an example post in the right-hand side panel.

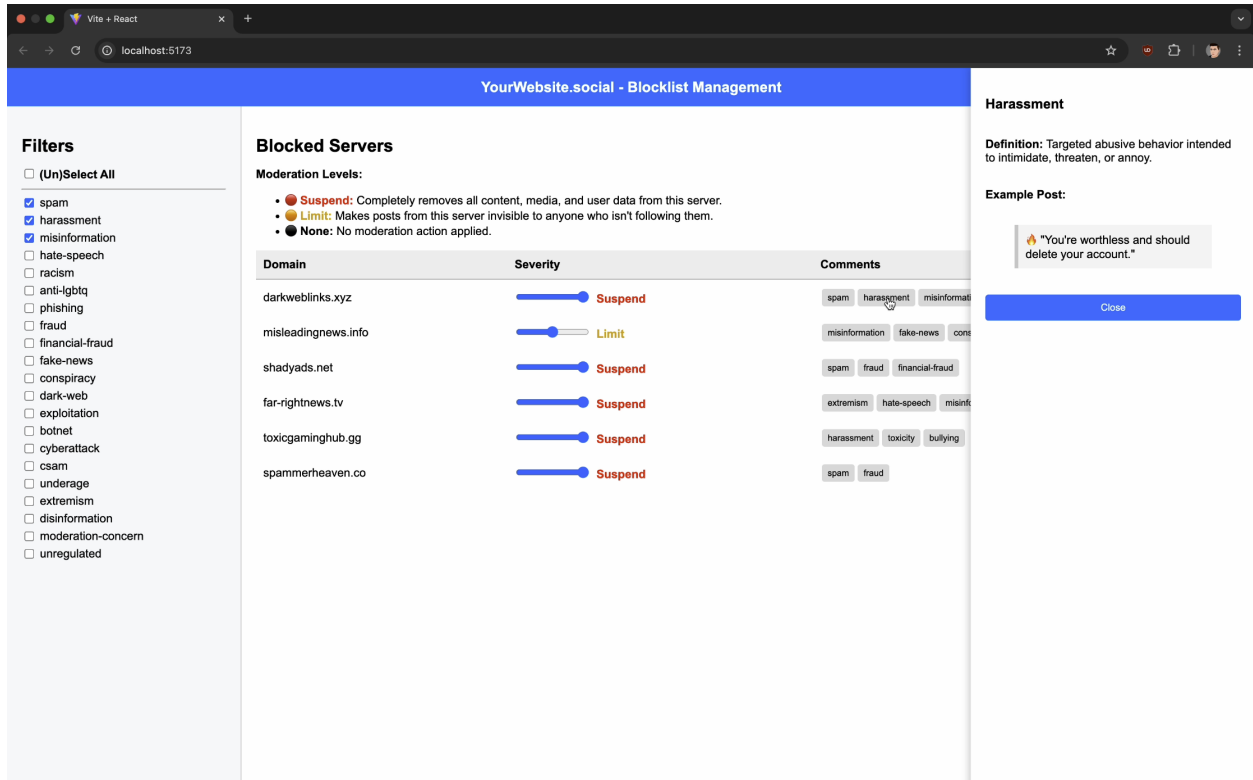


Figure 3.1: Screenshot of the demo system, showcasing category filters, severity toggles, and comments.

### 3.2.4 Findings for RQ3

To answer RQ3 (What design features would enhance the usability and effectiveness of community-level blocklist management tools?), we showed our demo system to participants and asked their opinions on the three features we designed, as well as what additional features they would like to have.

#### What do participants think about the demo system?

We break this into three parts based on their thoughts about: category filters, severity toggles, and comments.

**About category filters.** Participants mentioned that the filter has good usage, and they also suggested ideas for filtering instances by severity, blocking date, and the number of reports received from each instance.

Many participants (P1, P5, P6, P8, P9, P11, P12) found the category filters useful, noting that they can help moderators quickly identify issues like spam and harassment (P5, P8, P11, P12), reduce moderation workload (P5, P6, P11, P12), organize blocklists (P1, P6, P8), and improve visibility over why instances are blocked (P1, P9). P1 stated that the filtering part is “something quite difficult to do right now—there’s no tools to do so.” Additionally, participants expressed their need for expanded filtering features. P7 wanted severity filters and a search function: “I would want to be able to also filter the list by the current suspension severity, or by free text search on the server names, comment fields, etc.” P6 also mentioned that filtering by time, such as the date of blocking, would be helpful for deciding whether to revisit blocked instances. As P6 explained, “(If) I blocked this three years ago. Maybe we should see if it still applies.” P11 thought it would be good to link the report database to the filters so that users could filter based on the number of reports each instance received.

**About severity toggles.** Participants wanted more granular control when choosing severities, and they suggested that the system should also include features such as rejecting media and rejecting reports.

P2 described the difficulty of choosing severity for instances in gray areas and wanted more control beyond just “Suspend,” “Limit,” or “None.” P9 said:

“This (only having ‘Suspend,’ ‘Limit,’ and ‘None’) isn’t a blocklist problem, this is a federation model problem—it’s all or nothing. There’s ‘Limited,’ which is kind of in between, but not very effective. ... If you’re blocking an instance, you’re completely blocking an instance. That’s often what you want—but not always. There are cases where there’s a handful of people I’d like to allow in on an allowlist basis.” — P9, 7+ years of experience in the Fediverse

P9 also mentioned that it is unusual to change severity, noting that the only common change people talk about is upgrading a limit to a full block. P7 and P9 further pointed out that Mastodon’s built-in system currently includes more options, such as rejecting

media files and rejecting reports from an instance, without necessarily limiting or suspending it—features they suggested should be included in the tool.

**About comments.** Participants found the comment feature useful and suggested additional improvements, such as multi-language support, receipt documentation, and a universal vocabulary, while also raising concerns about trusting tags.

P2, P5, P6, and P8 found the comment feature very useful for documenting reasons for blocking decisions (P5, P6), improving transparency for both moderators and users (P2, P5), organizing and categorizing server actions more effectively (P8), and providing context for future moderation reviews (P2, P6).

Additionally, participants wanted further enhancements to the comment feature. P10, who manages a Chinese-speaking instance, argued that it would be helpful to have multi-language support for the custom tag function. P1, P7, P8, and P9 wanted a “receipt” feature. According to P1, the “receipt” would provide context—“when this person acts, what is the date, and who did what.” P7 suggested including more information beyond tags, such as the blocking date, blocking source (if the block was inherited from another blocklist), and free text comments. P9 proposed a shared public receipt library similar to Seirdy’s receipts and The Bad Space.

Regarding the content of comments, P1 recommended developing a universal comment vocabulary, stating that “a tag is useful if it’s recognized by the community.” On the other hand, P3 and P4 expressed skepticism about trusting tags. P3 felt it is difficult to characterize a group based on the behavior of one or two individuals, while P4 simply stated they “don’t have faith in tags.”

### **What features do they want?**

We found that participants expressed diverse expectations and needs regarding desired features. We organized their responses into four themes: (1) community collaboration features, (2) efficiency and management tools, (3) better monitoring and visibility, and (4)

customization and flexibility.

**Community collaboration features** Participants emphasized the importance of collective and collaborative approaches to blocklist management, including voting on instances, and sharing, syncing, selecting, and subscribing to blocklists. Within-instance collaboration was a common focus. P1 proposed a voting system to enhance internal moderator communication and decision-making: “Maybe you want to ask the team of moderators for their advice. And maybe we could have something like a voting system or a button to tell the team, ‘We have a question on this one. Can we discuss? Or maybe chat, or leave notes.’” Managing multiple instances, P5 expressed interest in the ability to share blocklists and automate blocking decisions across instances they moderate, “not publicly, just between our own instances.” For inter-instance collaboration, participants suggested leveraging collective signals. P5 proposed a public “suspect list” where moderators could comment and vote on instances’ status. P8 envisioned “scorecards” indicating how restrictive or permissive a blocklist is, explaining:

“For some servers who are very restrictive, ‘we have the safest block list’—you can see it here. And you don’t even have to show it to the world. You can just say, this is our block list score—for how restrictive or permissive it is. There are many servers that actively do not share their block list for safety reasons. ... What if you were to go to a server and it said, here’s our block list scorecard—we block 100% of CSAM servers. That’s a strong statement that would appeal to almost everybody.” - P8, owner of a local-based instance

P4 introduced Bluesky’s “starter pack” approach—allowing users to subscribe to curated lists—and recommended similar features for Mastodon, such as starter-pack blocklists sourced from trusted sources. Similarly, P8 suggested a “marketplace of blocklists,” where users could choose and subscribe to different blocklists tailored to their preferences.

**Efficiency and management tools** Participants also highlighted the need for tools that enhance efficiency and streamline moderation workflows. As a small instance owner, P1

emphasized the need for automated services capable of “finding every problematic server on the Fediverse,” stating willingness to pay for such a service because “it’s essential.” Ideally, such tools would be integrated into the default Mastodon interface—minimizing reliance on third-party services (P1). P4 echoed this, saying, “I’m sure something like this—if you could click on something and it automatically blocks it from your instance—would be helpful.” Participants also emphasized the importance of monitoring blocked instances over time. P2 reflected,

“Maybe someone thinks the mods aren’t doing a good job, but what if they change and start doing a good job after six months? How would you know that? [...] Especially since the action is pretty drastic, right? Like, ‘We’re going to block your entire instance and not give you a chance to improve or respond.’” — P2, using FediCheck

P3 proposed a mechanism for servers to petition for unblocking: “For example, it could show that an instance has petitioned for unblocking—so they’re going through the process of making that request.”

**Better monitoring and visibility** Participants discussed the need for richer, large-scale information to support moderation decisions. P1 emphasized several key metrics: whether an instance is active, the number of users on the instance, and the number of users from their own instance who follow it. These factors would significantly impact decisions—P1 noted they would avoid blocking inactive servers, treat large instances more seriously, and consider potential disruptions caused by users losing followers. P9 suggested incorporating information about whether other instances are also blocking a server, similar to Fediseer’s approach. As previously mentioned, the number of reports would be another important information, according to P2. Beyond these, P4 proposed adding instance-level descriptive statistics to inform decision-making:

“I think it would be helpful if there were a way to examine large-scale, instance-level trends—maybe something averaged across users—to get a better

sense of what’s going on at the domain level. [...] I’d want to know how many people are on a given domain, their post or comment frequency, and maybe some kind of quantitative measure that gives a high-level sense of the content.”

**Customization and flexibility** Participants emphasized the need for richer tools to support flexible, context-sensitive moderation decisions. Beyond basic blocking, they discussed features like early warnings, comparison tools, and accessibility improvements. P9 suggested a “pre-block” queue system to alert moderators when new instances begin appearing in shared blocklists or catalogs: “If they’re harassing other people, I want to block them before they start harassing me.” Rather than automatic blocking, P9 envisioned a watchlist that would allow proactive moderation. Similarly, P8 proposed features for comparing different blocklists, describing it as a way to reflect on moderation standards: “The things you block are shared across 50% of everyone else. That’s interesting. I’d start wondering—am I being too hard on some people?” Finally, P9 raised the importance of alt text and accessibility as integral to community values, envisioning tools that could support collaborative improvements for disabled users.

# Chapter 4

## Discussion

In this chapter, we reflect on our findings and discuss broader implications for moderation practices, community governance, and the design of future moderation tools. We distill these insights into three key themes.

### 4.1 Openness vs. Safety: The Tolerance of False Positives

Our findings in Chapter 3.2.2 revealed several moderation styles among Mastodon administrators. In this section, we group *exposure-first* and *report-and-fix* approaches as **openness-preferred** styles, and *safety-first* and *find-and-act* as **safety-preferred** styles. We observed a consistent pattern: openness-preferred moderation tends to be more reactive, relying on user reports and valuing exposure to diverse content, while safety-preferred styles adopt a more proactive stance, preemptively identifying and mitigating potential harms. This distinction aligns with prior work highlighting the dichotomy between reactive and proactive moderation paradigms Schluger et al. [2022]. The trade-off between openness and safety is a longstanding tension in social media moderation. Jiang et al. [2023] articulate this as a choice between “nurturing” and “punitive” moderation philosophies, reflecting a deeper normative conflict between fostering free expression and preventing harm. While this dilemma is well documented in traditional platforms, it becomes especially pronounced in *decentralized contexts*, where

governance is more decentralized and responsibility more distributed.

Decentralized platforms like Mastodon champion principles of *free speech* and *openness* Spencer-Smith [2025]. As Mansoux and Roscam Abbing Mansoux and Roscam Abbing [2020] note, federated social media projects extolled openness, universality, and free information flow, echoing liberal free-speech ideals. However, this same openness also introduces structural challenges. Roth and Lai Roth and Lai [2024a] describe how federation diffuses responsibility, making it hard to respond to malicious actors who can easily hop instances.

On the one hand, decentralized platforms prioritize more safety in their established rules. Nicholson et al. Nicholson et al. [2023a] find that rules prohibiting hate speech and harassment are significantly more prevalent in Mastodon instances than in Reddit communities. For example, 71% of Mastodon servers in their sample explicitly prohibit hate speech, compared to just 1.6% of subreddits. Similarly, rules on harassment (64% vs. 2.7%), spam (42% vs. 1.7%), and illegal content such as CSAM (48% vs. not reported) are far more common in Mastodon. These rules align closely with the top areas where Mastodon moderators report needing the most guidance.<sup>1</sup>

On the other hand, not all servers pursue "mass defederation"—a safety-first strategy that involves blocking large numbers of instances preemptively. Spencer-Smith and Tomaz Spencer-Smith and Tomaz [2025] find that most administrators continue to follow a *default federation* model, maintaining open communication with other instances and applying blocks only on a case-by-case basis:

"Most instances follow hashtags such as #fediblock or blocklists, but claim to still not adopt mass defederation, rather opting for default federation with all instances in the Fediverse, blocking unwanted instances on a case-by-case basis, in a practice that predates the influx of new users in 2022." - Spencer-Smith and Tomaz [2025]

---

<sup>1</sup><https://connect.iftas.org/wp-content/uploads/2024/04/IFTAS-Fediverse-Trust-and-Safety-Needs-Assessment-Report-Q3-2023.pdf>

The debate between case-by-case moderation and blanket interventions echoes broader platform dynamics. Jhaver Jhaver [2025] compares *community bans* (aligned with safety-preferred moderation) and *warning labels* (aligned with openness-preferred moderation), finding that users tend to favor measured, contextual interventions over one-size-fits-all enforcement. This resonates strongly with the *careful-decision* style we observed in our interviews, where moderators emphasized the responsibility and impact of blocking entire instances.

Ultimately, as Kissane and Kazemi observe Kissane and Kazemi [2024], “no single moderation style fits all.” Communities continuously navigate trade-offs between openness and safety, adjusting one pressure (e.g., strict rules or expanded moderator teams) to balance another (e.g., open registration or federation policies). This dynamic flexibility is a core feature of federated governance: rather than adhering to a fixed moderation model, instances evolve their practices based on local needs, values, and capacities.

## 4.2 Hero or Villain? Exploring Moderators’ Influence on Community Dynamics

In traditional social media, moderators play different roles such as **guardians** who remove harmful actors to protect their community members, and **gatekeepers** who set boundaries around allowed norms and content Seering et al. [2022]. Prior research indicates that these combined roles can lead to significant power concentration Matias [2019]. Within decentralized platforms, the role and influence of moderators become even more critical and amplified due to each instance independently setting its own moderation policies. Spencer-Smith and Tomaz Spencer-Smith [2025] found that Mastodon’s federated structure empowers instance administrators and moderators to unilaterally defederate (block) other servers perceived as violating their norms, effectively isolating entire communities. Our findings in Chapter 3.2.2 reinforce this perspective, showing that moderators are highly

cautious in decision-making, fully aware that their actions impact entire communities—they can be the heroes, and they can also be the villains.

**Moderators as Heroes.** A prominent example of moderators acting “heroically” is illustrated by the 2015 Reddit Blackout <sup>2</sup>, where subreddit moderators collectively set their communities to private to protest controversial platform policies. Matias Matias [2016] describes how, by temporarily withholding access, moderators effectively leveraged their collective gatekeeping power to advocate for improved governance and transparency from the platform. In this context, moderators positioned themselves as protectors, standing up for their communities’ interests, autonomy, and well-being.

**Moderators as Villains.** Conversely, moderation practices can lead to negative or unintended consequences. Halfaker et al. Halfaker et al. [2013] documented how Wikipedia’s implementation of stringent moderation tools—such as algorithmic reversion of edits and increased gatekeeping measures—unintentionally created an unwelcoming environment for newcomers. Despite intentions to safeguard content quality, these practices significantly reduced newcomer participation and diminished community vibrancy. Moderators, in their gatekeeping capacity, inadvertently created barriers rather than protection.

Moreover, community leaders themselves can become sources of division, exemplified by the cancellation of the decentralized community event, FediForum<sup>3</sup> in April 2025. The controversy arose from transphobic statements resurfaced from a former event co-organizer. Although moderators and administrators responded swiftly by withdrawing support and explicitly citing inclusivity concerns, the incident highlighted how influential figures within the community can inadvertently undermine cohesion.

Collectively, these examples underscore the significant and complex impact moderators, as community leaders, wield. We advocate for moderators aware of both the necessity of

---

<sup>2</sup><https://socialmediacollective.org/2015/07/09/what-just-happened-on-reddit-understanding-the-moderator-blackout/>

<sup>3</sup><https://wedistribute.org/2025/03/fediforum-canceled/>

moderation in safeguarding online communities and the inherent risks of concentrating power without robust accountability mechanisms.

## **4.3 Rethinking Blocklist Infrastructure—Designing for Collaborative and Transparent Moderation**

Our findings in Chapter 3.2.2 and Chapter 3.2.4 highlight moderators’ diverse perceptions and needs regarding community-level blocklists and blocklist-related tools. While some moderators value external blocklists for their efficiency, collective safety, and support for informed decision-making through shared resources, others express skepticism, raising concerns about biases, transparency, and the necessity of contextual decision-making. Through our content analysis in Chapter 3.1, we identified potential design improvements for future moderation tools, emphasizing transparency and community collaboration. In this section, we propose these two sets of design implications.

### **4.3.1 Supporting Collaborative Moderation Across Communities**

Prior work highlights moderation as a collective, community-driven process in which volunteer moderators collaboratively make governance decisions Matias [2019]. This is especially relevant in decentralized governance contexts such as Mastodon, where moderation heavily depends on instance-specific norms and practices. Jhaver et al. Jhaver et al. [2023] argue that moderation tools tailored to individual community standards empower local moderators by better aligning moderation processes with community values. Despite several third-party moderation tools already being applied within Mastodon, our findings suggest there is still significant demand for collaborative features explicitly designed to facilitate inter-instance communication and cooperation.

One critical area for supporting collaborative moderation involves the management and application of external blocklists. Although popular blocklists currently exist, participants

emphasized the need for improved mechanisms that facilitate discovering, comparing, subscribing to, and merging these blocklists—essentially envisioning a robust “blocklist marketplace.” Similar concepts have already been explored by platforms such as Bluesky through their “starter packs”, where users can easily preview curated lists and seamlessly apply moderation settings. However, Mastodon instances present a more complex context compared to individual user profiles, making it challenging to succinctly represent an entire instance’s moderation profile or reputation.

Structured documentation of moderation actions, such as moderation receipts, can help moderators clearly track decisions and foster accountability, as noted by prior research Geiger and Ribes [2011]. Designers should seriously consider enhancing moderation comments and receipt documentation by including richer contextual metadata: the involved moderators (actors), timestamps (timing), justification for actions, public shared receipt libraries (such as those provided by Seirdy’s Receipts or The Bad Space), universal vocabulary tags recognized across multiple instances, and multilingual support for global collaboration.

On decentralized platforms such as Mastodon, communities rely on shared norms and deliberate consensus-building processes to govern effectively. The concept of a “digital covenant” underscores the importance of establishing mutual agreements among community members to maintain and enforce moderation standards Zulli et al. [2023]. However, designers must carefully consider potential risks of manipulation and ensure mechanisms exist to identify and validate trustworthy sources. Consequently, moderation tools should incorporate community-driven voting or deliberation systems to collectively assess and signal the trustworthiness or reputation of instances, thereby providing transparent indicators that can guide community decision-making.

### 4.3.2 Enhancing Trustworthy Transparency in Moderation Systems

Transparency is a critical aspect of moderation governance, helping platforms establish legitimacy through openly justified actions Gillespie [2018], enhancing user trust Binns et al. [2018], and reducing perceptions of unfairness Epstein et al. [2022]. Prior work has propose shared safety mechanisms: e.g. federated trust networks among admins, transparent and interoperable moderation data, and possibly centralized support infrastructure for things like child safety Roth and Lai [2024a].

Our findings suggest that a key reason experienced moderators hesitate to adopt or trust external blocklists is the insufficient contextual information about how these blocklists are created and maintained. To address this issue, moderation tools can leverage clear, detailed analytics to enable moderators to make more informed decisions Jhaver et al. [2019a]. Specifically, tools should provide moderators with richer, domain-level metrics such as instance activity levels, user population size, moderation histories, number of reports received, and instance reputation scores. Additionally, moderation tools should explicitly present key metadata, including the blocklist sources, selection criteria, and potential biases or limitations, enabling moderators to critically evaluate external lists before adopting them.

Moreover, moderation transparency remains equally important after decisions have been made. Eslami et al. Eslami et al. [2018] demonstrated that clear visual explanations of system-generated decisions significantly improve user understanding and trust. Although moderation on Mastodon primarily involves manual decision-making, explicitly documenting moderation rationales similarly enhances trust among users, peer moderators, and the broader federated community. Moderators should be encouraged to communicate details such as selection criteria, reasons for blocking, and categories of blocked content. Such disclosures provide prospective users greater flexibility and context when selecting

instances to join. Ultimately, we propose that clearly documented moderation decisions and publicly available blocklists can serve as complementary resources to an instance's homepage and stated rules, collectively fostering an environment of informed participation and trust.

# Chapter 5

## Limitations

Our study has several limitations. First, the small sample size of 12 Mastodon moderators, while offering valuable insights, limits the generalizability of our findings across the broader moderator group. Second, our content analysis focused on a selection of publicly visible and widely referenced blocklists and tools, which may introduce selection bias and overlook less prominent resources. Third, the dynamic and evolving nature of moderation practices in decentralized networks poses challenges for capturing a complete and stable picture; practices and norms may shift over time as platforms grow and communities adapt. Finally, the decentralized structure of Mastodon fosters highly diverse, context-specific community standards, making it difficult to generalize findings or propose universally applicable moderation strategies across instances.

# Chapter 6

## Conclusion

Moderation practices such as community-level blocklists in Mastodon present new opportunities and challenges for community governance. Through content analysis and interviews with Mastodon moderators, our study highlights how community-level blocklists serve as important tools for maintaining community safety while reflecting diverse moderation philosophies and practices. We found that moderators engage with blocklists cautiously, balancing the need for proactive protection with concerns about overblocking and false positives. Our work identifies key opportunities for designing blocklist management tools that are flexible, transparent, and collaborative, ultimately supporting more resilient decentralized communities. As decentralized networks continue to evolve, understanding the trade-offs between openness and safety, as well as the significant responsibilities moderators hold in these emerging contexts, will be critical for sustaining healthy and inclusive online spaces.

# Bibliography

Ishaku Hassan Anaobi, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola, and Gareth Tyson. Will admins cope? decentralized moderation in the fediverse. In *Proceedings of the ACM Web Conference 2023*, pages 3109–3120, 2023.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. ‘it’s reducing a human being to a percentage’: Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

Carlo Bono, Lucio La Cava, Luca Luceri, and Francesco Pierri. An Exploration of Decentralized Moderation on Mastodon. In *Proceedings of the 16th ACM Web Science Conference (WebSci ’24)*, page 6, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3614419.3644016.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction*, 1(CSCW):1–22, 2017.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet’s hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

Kate Crawford and Tarleton Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *new media & society*, 18(3):410–428, 2016.

Anwitaman Datta, Sonja Buchegger, Le-Hung Vu, Thorsten Strufe, and Krzysztof Rządca.

- Decentralized online social networks. *Handbook of social network technologies and applications*, pages 349–378, 2010. doi: [https://doi.org/10.1007/978-1-4419-7142-5\\_17](https://doi.org/10.1007/978-1-4419-7142-5_17).
- Ziv Epstein, Sydney Levine, David G. Rand, and Iyad Rahwan. Do explanation modalities make ai decisions fairer? an empirical investigation of human perceptions of ai fairness. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.
- Motahhare Eslami, Siddharth Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Algorithmic transparency in the age of ai. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14. ACM, 2018.
- Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.
- R. Stuart Geiger. Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6):787–803, 2016. doi: 10.1080/1369118X.2016.1153700.
- R Stuart Geiger and David Ribes. Trace ethnography: Following coordination through documentary practices. *Proceedings of HICSS 2011*, pages 1–10, 2011.
- Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.
- W3C Social Web Working Group. Activitypub. <https://www.w3.org/TR/activitypub/>, 2018. W3C Recommendation.
- Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. The rise and decline of an open collaboration system: How wikipedia’s reaction to popularity is

causing its decline. *American Behavioral Scientist*, 57(5):664–688, 2013. doi: 10.1177/0002764212469365.

Anaobi I. Hassan, Aravindh Raman, Ignacio Castro, Haris B. Zia, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. Exploring Content Moderation in the Decentralised Web: The Pleroma Case. In *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies (CoNEXT '21)*, pages 328–335, New York, NY, USA, 2021a. Association for Computing Machinery. doi: 10.1145/3485983.3494838.

Anaobi Ishaku Hassan, Aravindh Raman, Ignacio Castro, and Gareth Tyson. The impact of capitol hill on pleroma: the case for decentralised moderation. In *Proceedings of the CoNEXT Student Workshop*, pages 1–2, 2021b.

Sohyeon Hwang, Charles Kiene, Serene Ong, and Aaron Shaw. Adopting third-party bots for managing online communities. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–26, 2024.

Sohyeon Hwang, Priyanka Nanayakkara, and Yan Shvartzshnaider. Trust and friction: Negotiating how information flows through decentralized social media. *arXiv preprint arXiv:2503.02150*, 2025.

Shagun Jhaver. Bans vs. warning labels: Examining bystanders’ support for community-wide moderation interventions. *ACM Trans. Comput.-Hum. Interact.*, 2025. doi: 10.1145/3715116. Early Access.

Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.*, 25(2):1–33, March 2018. doi: 10.1145/3185593.

Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine

- collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019a.
- Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019b.
- Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X Zhang. Designing word filter tools for creator-led comment moderation. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–21, 2022.
- Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33, 2023.
- J. A. Jiang, P. Nie, J. R. Brubaker, and C. Fiesler. A trade-off-centered framework of content moderation. *ACM Trans. Comput.-Hum. Interact.*, 30(1):3:1–3:35, 2023. doi: 10.1145/3534929.
- Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design*, 1:4–2, 2012.
- Erin Kissane and Darius Kazemi. Seven themes in fediverse governance: Findings report on fediverse microblogging servers. <https://fediverse-governance.github.io/>, 2024.
- Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*, pages 933–943, 2018.

- Yuhan Liu, Emmy Song, Owen Xingjian Zhang, Jewel Merriman, Lei Zhang, and Andrés Monroy-Hernández. Understanding decentralized social feed curation on mastodon. *arXiv preprint arXiv:2504.18817*, 2025.
- Kaitlin Mahar, Amy X Zhang, and David Karger. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.
- Aymeric Mansoux and Roel Roscam Abbing. Seven theses on the fediverse and the becoming of floss. In *The Eternal Network: The Ends and Becomings of Network Culture*, pages 124–140. Institute of Network Cultures and Transmediale, 2020.
- Mike Masnick. Protocols, not platforms: A technological approach to free speech. <https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech>, 2021. Knight First Amendment Institute at Columbia University.
- J. Nathan Matias. Going dark: Social factors in collective action against platform operators in the reddit blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1138–1151. ACM, 2016. doi: 10.1145/2858036.2858391.
- J. Nathan Matias. The civic labor of volunteer moderators online. *Social Media + Society*, 5(2):1–12, 2019. doi: 10.1177/2056305119836778.
- Aiden McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Cranefield. Controlling bad behavior in online communities: An examination of moderation work. 2016.
- Matthew N. Nicholson, Brian C. Keegan, and Casey Fiesler. Mastodon rules: Characterizing formal rules on popular mastodon instances. In *Companion Publication of the 2023 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW Companion)*, pages 103–107, 2023a. doi: 10.1145/3584931.3606970.

- Matthew N. Nicholson, Brian C. Keegan, and Casey Fiesler. Mastodon Rules: Characterizing Formal Rules on Popular Mastodon Instances. In *Companion of the 2023 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '23 Companion)*, page 5, New York, NY, USA, 2023b. Association for Computing Machinery. doi: 10.1145/3584931.3606970.
- Karl Popper, Ernst Hans Gombrich, and Vaclav Havel. *The open society and its enemies*. Routledge, 2012.
- Sarah T. Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019.
- Eugen Rochko. Mastodon. <https://joinmastodon.org>, 2016. Decentralized social network software.
- Yoel Roth and Samantha Lai. Securing federated platforms: Collective risks and responses. *Journal of Online Trust and Safety*, 2(2):1–28, 2024a. doi: 10.54501/jots.v2i2.171.
- Yoel Roth and Samantha Lai. Securing federated platforms: Collective risks and responses. *Journal of Online Trust and Safety*, 2(2):1–22, 2024b. doi: 10.54501/jots.v2i2.171.
- Alan Z. Rozenshtein. Moderating the Fediverse: Content Moderation on Distributed Social Media. *Journal of Free Speech Law*, 3:217–236, 2023.
- Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. Proactive moderation of online discussions: Existing practices and the potential for algorithmic support. *Proc. ACM Hum.-Comput. Interact. (CSCW)*, 6(CSCW2):370:1–370:29, 2022. doi: 10.1145/3555095.
- Joseph Seering, Robert Kraut, and Laura Dabbish. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM*

*conference on computer supported cooperative work and social computing*, pages 111–125, 2017.

Joseph Seering, Geoff Kaufman, and Stevie Chancellor. Metaphors in moderation. *New Media & Society*, 24(3):621–640, 2022. doi: 10.1177/1461444820964968.

Charlotte Spencer-Smith. Labour pains: Content moderation challenges in mastodon growth. *Internet Policy Review*, 14(1):1–21, 2025.

Charlotte Spencer-Smith and Tales Tomaz. Labour pains: Content moderation challenges in mastodon growth. *Internet Policy Review*, 14(1), 2025. doi: 10.14763/2025.1.1831.

Jing Zeng and D Bondy Valdovinos Kaye. From content moderation to visibility moderation: A case study of platform governance on tiktok. *Policy & Internet*, 14(1): 79–95, 2022.

Cindy C Zhang, Grayden Zaleski, Jaya N Kailley, Katelyn A Teng, Mahala English, Anna Riminchan, and Julie M Robillard. Debate: Social media content moderation may do more harm than good for youth mental health. *Child and Adolescent Mental Health*, 29(1):104–106, 2024.

Diana Zulli, Timothy Liu, and Payal Arora. The digital covenant: Non-centralized platform governance on the mastodon social network. *New Media & Society*, 25(2):309–327, 2023.

# Appendix A

## Full Table of Content Analysis

A.1 Blocklists Analysis (Tables A.1 and A.2)

A.2 Blocklist-related Tools Analysis (Tables A.3 and A.4)

| Category                            | Garden Fence                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                              | CARIAD                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | IFTAS-DNI                                                                                                         |                                                                                                                                                                |
|-------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                     | Summary                                                                                                                                                                                                                                                                                                             | Content                                                                                                                                                                                                                                                                                                                                                                                                      | Summary                                                                                                                                                                                                                   | Content                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Summary                                                                                                           | Content                                                                                                                                                        |
| Purpose                             | Garden Fence is a starter blocklist for Mastodon admins to protect users from well-known sources of harmful content.                                                                                                                                                                                                | Garden Fence serves as a starting point for Mastodon server admins seeking to protect their users from the worst and most well-known sources of harmful content, such as hate speech, racism, sexism, and similar forms of abuse.                                                                                                                                                                            | CARIAD provides new moderators a basic, first step in blocking the most widely blocked domains.                                                                                                                           | CARIAD is intended to provide new admins a basic, first step to block the most widely blocked domains, since people(IFTAS survey respondents, community members, advisors) asked for a shared knowledge of the most commonly blocked domains.                                                                                                                                                                                                                        | DNI (Do Not Interact) List is a subset of CARIAD, containing domains that are highly recommended to block.        | DNI (Do Not Interact) List is an IFTAS-curated list of domains that are highly recommended for defederation.                                                   |
| Criteria                            | Garden Fence blocks domains listed by sunny.garden and most reference servers, but excludes those that are offline or outside its intended scope, even if widely blocked.                                                                                                                                           | Garden Fence includes a domain if it is blocked by sunny.garden and at least 6 out of 7 additional reference servers. Garden Fence excludes domains that are believed to be offline or outside the intended scope of the blocklist, even if they are widely blocked—for example, threads.net, which is commonly blocked but falls outside Garden Fence's focus.                                              | CARIAD compiles data from the IFTAS DNI list and from domain blocks used by large, established Mastodon service providers that meet specific eligibility criteria.                                                        | CARIAD combines data from two sources: the IFTAS DNI List, a manually reviewed list of domains labelled by harm., and a curated aggregation of the blocks in place on high volume Mastodon service providers that meet criteria including service duration, user base size, existing block practices, and exclusion from the CARIAD or DNI lists.                                                                                                                    | DNI applies labels sourced from the DTSP Trust & Safety Glossary to categorize the actors, behavior, and content. | DNI uses the following labels and definitions to label actors, behaviour, and content. Labels are sourced from the Trust & Safety Glossary of Terms from DTSP. |
| Transparency                        | Garden Fence manually reviews each server, selects reference servers based on shared values and independent operation; however, servers and selection criteria are not publicly disclosed to minimize administrative overhead. The blocklist includes public comments explaining the reasons for each domain block. | Garden Fence made all updates after manual reviews. Reference servers, where they use to check for blocks, are carefully selected based on compatible values and independent operation, but are not publicly listed and the selecting criteria is not openly shared to reduce administrative overhead. The blocklist includes public comments to explain why a domain was blocked.                           | CARIAD includes domains blocked by at least 51% of observed sources or listed on the list, with each domain reviewed by IFTAS staff or advisors before inclusion, and offers an email-based appeal process for delisting. | CARIAD includes domains blocked by at least 51% of observed sources or listed on the IFTAS DNI list. Each domain undergoes manual review by IFTAS staff or advisors before inclusion. The database is reviewed quarterly to ensure fairness, removing domains that no longer meet criteria. Appeals can be made via email for IFTAS review. During a 60-day pilot, three domains were added: two for illegal content and one for spam.                               | NA                                                                                                                | NA                                                                                                                                                             |
| Distribution                        | Garden Fence is publicly available and updated weekly.                                                                                                                                                                                                                                                              | Garden Fence csv files are updated and reviewed weekly: one allows direct import to Mastodon, one is compatible with fediblock-sync tool.                                                                                                                                                                                                                                                                    | CARIAD is available at tool Fedicheck.                                                                                                                                                                                    | CARIAD is free and available via the FediCheck service.                                                                                                                                                                                                                                                                                                                                                                                                              | DNI is publicly available online.                                                                                 | DNI is publicly available on IFTAS Connect library webpage.                                                                                                    |
| Limitations                         | Garden Fence is curated based on sunny.garden's judgments, reflecting specific biases and values, primarily focusing on English-language content.                                                                                                                                                                   | Garden Fence is curated based on the judgments of sunny.garden and is not neutral or unbiased. It also does not necessarily represent all of the different kinds of unwanted content depending on the goals and values of particular servers, such as porn, crypto, bots, bridges, etc. Also, it focuses primarily on English-language content and may not cover content in other languages comprehensively. | CARIAD is biased toward global north perspectives, lacks protection for marginalized communities, and is only a short-term starting point for domain curation.                                                            | CARIAD's inclusion criteria, favoring North American, Western European, and Southeast Asian service providers, introduces a bias toward perspectives from the Global North and white-majority regions. And CARIAD does not protect LGBTQ, BIPOC, BAME or other marginalised communities. Additionally, it is highly recommended that new service providers first use CARIAD to <b>begin</b> their domain curation activities, and then explore additional resources. | NA                                                                                                                | NA                                                                                                                                                             |
| Other Resources (Blocklists, Tools) | Garden Fence recommends additional blocklists and tools.                                                                                                                                                                                                                                                            | Garden Fence website recommended several other blocklists like Seirdy's one, and Awesome Fedi Admin Resources that has a list of Basic Blocklists.                                                                                                                                                                                                                                                           | CARIAD recommends tool resources for synchronizing and managing Mastodon admin domain blocks.                                                                                                                             | CARIAD recommended to consult resources such as FediBlockHole, a tool for synchronizing Mastodon admin domain blocks; Mastodon Block Tools, a compilation of various block management projects; and guides on importing pre-made server blocklists.                                                                                                                                                                                                                  | NA                                                                                                                | NA                                                                                                                                                             |
| Additional Content                  | NA                                                                                                                                                                                                                                                                                                                  | NA                                                                                                                                                                                                                                                                                                                                                                                                           | CARIAD invites public feedback and offers channels for Mastodon administrators to communicate.                                                                                                                            | CARIAD is open for public feedback by email, IFTAS Connect Feedback forum, matrix group chat. It also offers group chats and channels for Mastodon administrators to collaborate and share best practices.                                                                                                                                                                                                                                                           | NA                                                                                                                | NA                                                                                                                                                             |

Table A.1: Content analysis of Mastodon blocklists: Garden Fence, CARIAD, and IFTAS-DNI.

| Category           | Seirdy Tier-0                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | FediNuke                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|--------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                    | Summary                                                                                                                                                                                                                                      | Content                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Summary                                                                                                                                                                                                           | Content                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Purpose            | Tier-0 is a starter blocklist that aligns with Mastodon Covenant standards and is meant to be adjusted by instance admins.                                                                                                                   | Tier-0 is designed as a starter blocklist to let new admins easily customize, following Mastodon Covenant moderation standards. It is a smaller, semi-curated subset of <a href="https://pleroma.envs.net">pleroma.envs.net</a> .                                                                                                                                                                                                                                                                                                    | FediNuke is used as a recommended default blocklist highlighting the most egregious servers.                                                                                                                      | FediNuke aims for broad defederation of the worst actors, characterized by severity, notoriety, and potential harm to other instances, without causing significant controversy.                                                                                                                                                                                                                                                                                     |
| Criteria           | Tier-0 is based on consensus across multiple blocklists.                                                                                                                                                                                     | Tier-0 is a subset of <a href="https://pleroma.envs.net">pleroma.envs.net</a> blocklist. It includes entries appearing on $\geq 15$ of 27 other instance blocklists (“bias sources”). It is built through consensus, but severity varies.                                                                                                                                                                                                                                                                                            | FediNuke is based on severity, blocking the most severe offenders.                                                                                                                                                | FediNuke is a curated subset of <a href="#">tier0.csv</a> , as the “worse half”. It is recommended as default blocklist for most users. It includes severe, notorious, or widely harmful instances, but excludes many bad but isolated instances to stay small. It is built through severity, but avoids controversial entries.                                                                                                                                     |
| Transparency       | Tier-0 is curated by an author who takes full accountability for its contents. The author selects reliable bias sources with proactive moderation, and occasionally overrides vote criteria based on severity, staff overlap, or clear risk. | Tier-0 has following process:<br>Seirdy, the author who is solely responsible for it, selects reference source, called “Bias Sources”, which must have accessible proactively moderated blocklists that cover $\geq 50\%$ of FediNuke entries. Current bias sources are not published.<br>Seirdy also applies overrides: instances with fewer than 15 votes might be included based on severity, staff overlap, or admin consensus for significant risks or clear bigotry; controversial but less harmful instances may be excluded. | FediNuke is a manually curated, severity-focused subset of Tier-0, supported by strict receipt criteria emphasizing staff involvement, clear harmful patterns, proper context, and third-party archived evidence. | FediNuke is a smaller, severity-based subset of Tier-0, manually curated to include only the most severe and notorious instances.<br>FediNuke receipt criteria are stricter, emphasizing multiple receipts per instance, clear evidence of staff endorsement or widespread problematic behavior, and relevant reblogs only from staff members boosting local content.<br>Each FediNuke entry includes a brief description and links to the original archived posts. |
| Distribution       | Tier-0 is publicly distributed in CSV form with an additional list to handle removals.                                                                                                                                                       | Tier-0 is publicly available on Seirdy’s website and CC-BY-SA licensed. It is updated manually and periodically after reviewing new entries; no fixed schedule specified.                                                                                                                                                                                                                                                                                                                                                            | FediNuke is publicly distributed in CSV form with less frequent updates and curated for severity.                                                                                                                 | FediNuke is publicly available on Seirdy’s website and CC-BY-SA licensed. It is update less frequently than Tier-0, as updates are manual and selectively curated for severity.                                                                                                                                                                                                                                                                                     |
| Limitations        | Tier-0 has limitations on transparency (hidden sources), objectivity (personal judgments), and timeliness (slow updates and corrections).                                                                                                    | Tier-0 has limitations on:<br>Subjectivity and personal bias: Entries can include overrides below consensus thresholds, based on the author’s judgment.<br>Not fully transparent: Bias sources (the curated instance blocklists used to form consensus) are not published, limiting external verification.<br>Delayed updates and retractions: Manual refreshes mean updates are infrequent, and retractions require a separate process.                                                                                             | FediNuke inherits Tier-0’s limitations and adds its own: limited completeness, minimal context, and slow updates.                                                                                                 | FediNuke shares Tier-0’s all limitations, plus:<br>Not comprehensive by design: Intentionally excludes many problematic instances if they appear less harmful or require extensive context.<br>Limited context for certain entries: Entries require minimal context; instances that demand detailed explanations may be excluded.<br>Slow and manual updates: Receipts and entries require multiple verifications, causing infrequent updates.                      |
| Other Resources    | Tier-0’s author, Seirdy, recommends FediBlockHole for managing blocklists and retractions, and also provides supplementary lists.                                                                                                            | Tier-0’s author recommends:<br>FediBlockHole: Recommended for Fediverse servers (currently Mastodon) to manage blocklists and support retractions effectively.<br>Seirdy maintains separate blocklists for specific moderation scenarios, including Twitter mirrors, social media bridges with weak moderation, and spammy or short-lived subdomains.                                                                                                                                                                                | Same with Tier-0                                                                                                                                                                                                  | Same with Tier-0                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Additional Content | Tier-0’s author, Seirdy, contributes beyond blocklists by documenting detailed moderation evidence (“receipts”), manually reviewing and updating lists, and actively responding to community feedback.                                       | Tier-0’s author collect and publish detailed “receipts” (documented evidence with archived links) explaining why instances are blocked. Seirdy also curates multiple blocklists ( <a href="https://pleroma.envs.net">pleroma.envs.net</a> , Tier-0, FediNuke) with varying severity and scope, including manual review and selective overrides. Seirdy also engages in ongoing moderation activities, including periodic manual updates, retractions, and adjustments based on community feedback.                                   | Same with Tier-0                                                                                                                                                                                                  | Same with Tier-0                                                                                                                                                                                                                                                                                                                                                                                                                                                    |

Table A.2: Content analysis of Mastodon blocklists: Seirdy Tier-0 and FediNuke blocklists.

| Category          | FediCheck                                                                                                                                |                                                                                                                                                                                                                                                           | Fediseer                                                                                                                                                |                                                                                                                                                                                                                                                                                                         |
|-------------------|------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                   | Summary                                                                                                                                  | Content                                                                                                                                                                                                                                                   | Summary                                                                                                                                                 | Content                                                                                                                                                                                                                                                                                                 |
| Summary           | Fecicheck is a service run by IFTAS that can auto-sync blocklists.                                                                       | FediCheck is developed and run by the IFTAS team as a service that automatically synchronizes Mastodon servers with curated blocklists, sparing admins from manual research and block entries.                                                            | Fediseer is a service created by developer Rikudou and maintained by db0 that provide spam/ham classification.                                          | Fediseer is a service for the fediverseto provide spam/harrassment classification, created by @rikudou@lemmings.world, and maintained by @db0@lemmy.dbzer0.com.                                                                                                                                         |
| Purpose           | FediCheck offers “moderation-as-a-service” by auto-applying a vetted denylist to help new Mastodon admins quickly set up block settings. | FediCheck focuses on providing a “moderation-as-a-service” tool for new Mastodon admins – admins log in and the service auto-applies a vetted denylist (and retractions) to jump-start their instance’s block settings.                                   | Fediseer uses a human-curated “chain of trust” system to classify Fediverse instances as spam or not and publicly record their approval or disapproval. | Fediseer provides a crowd-sourced human-curated spam/ham classification of Fediverse instances, and offers a public space to specify approval/disapproval of other instances. Through a human-driven system called the “chain of trust”, fediseer can help to figure out if an instance is spam or not. |
| Technical Access  | FediCheck is a web app that lets Mastodon admins (v4.1+) sign in and remotely apply or remove domain blocks on their server.             | FediCheck is accessed through a web application where an admin signs in with their server (requires Mastodon v4.1+ API), after which it can remotely apply or remove domain blocks on that server.                                                        | Fediseer offers a machine-readable API to access its spam classification.                                                                               | Fediseer provides a machine readable API to consume the data contained within.                                                                                                                                                                                                                          |
| Decision Type     | FediCheck applies CARIAD blocklist.                                                                                                      | FediCheck automates instance-level block decisions (suspensions or silences) by applying a trusted aggregated blocklist (combining IFTAS’s “CARIAD” and “Do Not Interact” lists), while ensuring each domain on the list has been reviewed by moderators. | Fediseer records three types of moderation judgements related to spam classification.                                                                   | Fediseer records three types of moderation judgements: guarantees (instance is trusted not-spam), endorsements (positive approval), and censures (negative flags).                                                                                                                                      |
| Tags / Categories | Fedicheck uses standardized labels to classify online behavior and content, with definitions and examples.                               | Fedicheck use labels from the DTSP(Digital Trust & Safety Partnership) Trust & Safety Glossary to classify actors, behavior, and content, each with definitions, guidance, and example community rules.                                                   | Fediseer allows customized up tp 100 descriptive tags for each instance. Moderators can use tags to filter their blocklists.                            | Fediseer allows instance admins to add up to 100 descriptive tags for their instance, which others can use to filter block/allow lists or to discover instances by category.                                                                                                                            |

Table A.3: Content Analysis of Mastodon blokcklist-related tools: FediCheck and Fediseer.

| Category          | FediBlockHole                                                                                                                                           |                                                                                                                                                                                                                                                                                    | The Bad Space                                                                                                                                                                                             |                                                                                                                                                                                                                                        |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                   | Summary                                                                                                                                                 | Content                                                                                                                                                                                                                                                                            | Summary                                                                                                                                                                                                   | Content                                                                                                                                                                                                                                |
| Summary           | FediBlockHole is an open-source tool that automatically syncs a Mastodon server's blocklist with trusted source servers.                                | FediBlockHole, created by github user eigenmagic, is an open-source tool (available on GitHub under user eigenmagic) that automatically syncs a Mastodon server's domain blocklist with blocklists from other trusted source servers.                                              | The Bad Space is a community-driven catalog of harmful instances with input from marginalized communities                                                                                                 | The Bad Space is a catalog of harmful instances, created by Ro and based on the #fediblock movement started by Marcia X and Ginger, with block data contributed by marginalized communities like rage.love and queer.group.            |
| Purpose           | FediBlockHole aggregates blocklists from trusted sources into one list and pushes it to Mastodon servers to simplify and coordinate moderation efforts. | FediBlockHole is intended to aggregate multiple blocklists from chosen trusted sources into one combined list and then push this unified blocklist out to one or more Mastodon instances, streamlining the sharing of block rules.                                                 | The Bad Space lists Fediverse servers that pose risks to marginalized users by hosting harmful content, bad actors, or lacking proper moderation.                                                         | The Bad Space exists to identify and list Fediverse servers that harbor bad actors, lack proper moderation, or host content like hate speech or CSAM – essentially, servers that present risks to marginalized users' safety.          |
| Technical Access  | FediBlockHole runs as a service that uses the Mastodon Admin API to fetch, import, and push blocklists between servers and sources.                     | FediBlockHole runs as a service or script that interfaces with the Mastodon Admin API – it can fetch blocklists from other Mastodon servers (via API, with or without admin auth as needed) and also import from URLs/files, then output or push updates via API to local servers. | The Bad Space offers a web-based searchable directory and a public API to share its list of flagged Fediverse instances with other tools and servers.                                                     | The Bad Space provides a web-based searchable directory of flagged instances and also offers a public API so that other moderation tools or servers can fetch its compiled lists of problematic instances.                             |
| Decision Type     | FediBlockHole lets instance admins subscribe to external blocklists while keeping final control over which domains stay blocked.                        | FediBlockHole allows instance admins choose which external blocklist maintainers to subscribe to, thereby delegating some blocking decisions, but the admin retains final control over which domains remain blocked on their server.                                               | The Bad Space lists an instance only after two or more independent communities have blocked it, preventing any single admin from making unilateral listing decisions.                                     | The Bad Space only lists an instance after two or more independent communities (its “Current Sources”) have decided to suspend or silence that instance, ensuring that no single admin's judgment alone puts a server on the list.     |
| Tags / Categories | FediBlockHole does not implement its own tagging or categorization system.                                                                              | FediBlockHole supports public and private comments on block entries but does not implement its own tagging or categorization system beyond Mastodon's standard fields.                                                                                                             | The Bad Space categorizes block reasons by harmful behavior types, focusing on instances promoting racism, sexism, transphobia, ableism, casteism, and similar content as flagged by partner communities. | The Bad Space categorizes the reasons for blocks by type of harmful behavior – it specifically targets instances promoting things like racism, sexism, transphobia, ableism, casteism, etc., as identified by its partner communities. |

Table A.4: Content Analysis of Mastodon blocklist-related tools: FediBlockHole and The Bad Space blocklist.