Attitudes, Imagined Roles, and Governance Boundaries for AI in Decentralized Social Media

Zhilin Zhang
zhilin.zhang@cs.ox.ac.uk
University of Oxford
United Kingdom

Jun Zhao
jun.zhao@cs.ox.ac.uk
University of Oxford
United Kingdom

Ge Wang
wangge@illinois.edu
University of Illinois
Urbana-Champaign
United States

Sruthi Viswanathan
sruthi.viswanathan@cs.ox.ac.uk
University of Oxford
United Kingdom

Tala Ross
tala.ross@cs.ox.ac.uk
University of Oxford
United Kingdom

Samantha-Kaye Johnston
samantha-kaye.johnston@cs.ox.ac.uk
University of Oxford
United Kingdom

Diyi Liu
diyi.liu@oii.ox.ax.uk
University of Oxford
United Kingdom

Hayoun Noh
hayoun.noh@cs.ox.ac.uk
University of Oxford
United Kingdom

Max Van Kleek
max.van.kleek@cs.ox.ac.uk
University of Oxford
United Kingdom

Nigel Shadbolt
nigel.shadbolt@cs.ox.ac.uk
University of Oxford
United Kingdom

Abstract

Decentralised social media (DSM) platforms such as Mastodon offer community-governed alternatives to corporate social networks but place substantial governance burdens on volunteer operators. As interest grows in applying artificial intelligence (AI) to support this work, little is known about whether DSM operators want AI, what roles they consider appropriate, and what governance boundaries they require. We conducted semi-structured interviews with 20 operators across Mastodon, Pixelfed, PeerTube, Lemmy, Pleroma, and Funkwhale, using generative feature probes and speculative scenarios to explore their perceptions of AI. Operators rejected AI as an autonomous actor, instead envisioning it as governance infrastructure that provides contextual intelligence, supports cross-instance coordination, and sustains community and moderator well-being. They also articulated strict boundaries rooted in DSM values, including human accountability, reversibility, transparency, community-centred configuration, and strong data-governance constraints. We contribute empirical insights and design implications for AI compatible with decentralised, federated social media.

CCS Concepts

• **Human-centered computing → Empirical studies in collaborative and social computing**.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790295

Keywords

decentralized social media, artificial intelligence, citizen-powered internet, admin experiences, moderation and governance, online communities

ACM Reference Format:
Zhilin Zhang, Jun Zhao, Ge Wang, Sruthi Viswanathan, Tala Ross, Samantha-Kaye Johnston, Diyi Liu, Hayoun Noh, Max Van Kleek, and Nigel Shadbolt. 2026. Attitudes, Imagined Roles, and Governance Boundaries for AI in Decentralized Social Media. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3772318.3790295

1 Introduction

Decentralised social media (DSM) platforms such as Mastodon have become important alternatives to corporate social networks whose centralised moderation, data extraction, and opaque algorithms have long produced inequitable outcomes — particularly for marginalised communities [38, 44, 47, 111]. DSM redistributes authority across independently operated servers ("instance"), enabling communities to set their own norms and data practices [1, 108, 112]. This autonomy is a key reason communities migrate to DSM, yet it also introduces substantial governance burdens: volunteer operators must manage conflict resolution, rule enforcement, and cross-instance harms such as harassment, misinformation, malicious actors, and child sexual abuse material (CSAM) [4, 95, 100]. These pressures pose sustainability challenges for an ecosystem built on volunteer labour.

Recent developments in artificial intelligence (AI) have significantly advanced the accuracy and depth with which automated

systems can analyse human-generated content, across a variety of application areas. In this paper, we examine whether automated or semi-automated tools could support governance work in DSM. While many forms of AI-based moderation are already used widely on centralised platforms [18, 39], studies of such systems have revealed harms resulting from biases and limitations in such AI systems, disproportionately impact marginalised users [28, 47, 87]. Moreover, the way these platforms are deployed is opaque, and centralised – applying unified policies, centralised data access, and hierarchical decision structures that do not exist in DSM. Technical explorations have begun to test whether machine-learning models can operate in decentralised settings — for example, using federated learning [109] or server-local classification [61] — but we still know little about whether DSM administrators and moderators (we use the term "operators" to collectively refer to these two roles) believe AI has a place in their communities, what roles they consider appropriate, and what governance boundaries they require.

This gap matters because DSM is not merely a smaller or slower version of centralised platforms. It is autonomous, heterogeneous, value-sensitive, and relationally governed. Harms frequently originate between instances rather than within them [53, 107]. Whether AI can be integrated meaningfully therefore hinges on the visions and boundaries articulated by those who sustain these environments.

We frame our investigation around these research questions:

- **RQ1**: What attitudes, views, and concerns do decentralised social media (DSM) operators have towards the use of AI in DSM?
- **RQ2**: What are the key areas of need where AI could positively support DSM, and what potential roles might it play in that support?
- **RQ3**: Can challenges and concerns on the use of AI in DSM be addressed through new governance mechanisms?

To examine these questions, we conducted an exploratory study informed by Research through Design (RtD) with 20 operators from Mastodon, Pixelfed, PeerTube, Lemmy, Pleroma, and Funkwhale. Through semi-structured interviews, generative feature probes, and speculative scenario probes [37, 79, 86, 110], we investigated operators' attitudes toward AI, the roles they envision for it, and the governance mechanisms they consider necessary. We used moderation as a situated entry point because it is a labour-intensive, high-stakes, and value-sensitive part of DSM governance [4, 95, 108] that naturally surfaces broader imaginaries and reflections of AI's potentials and risks.

Our findings show that DSM operators do not imagine AI as an autonomous decision-maker. Instead, they envision AI as governance infrastructure that: (1) provides contextual intelligence for informed, situated judgment; (2) supports coordination and early-warning functions across federated instances; and (3) helps sustain community and moderator well-being. At the same time, they draw strict governance boundaries grounded in DSM values: human accountability, reversibility, transparency, community-centred configuration, and strong data-governance constraints.

Our contributions are threefold: (1) we offer the first in-depth qualitative account of how DSM operators imagine and evaluate AI within DSM governance; (2) we surface the potential roles and areas of need operators believe AI could meaningfully support; and (3) we identify governance considerations — centred on autonomy, transparency, and data boundaries — that must shape any responsible AI integration in DSM.

2 Background

2.1 Decentralised Social Media Governance and Moderation

Decentralised social media (DSM) redistribute authority across independently operated servers ("instances"), contrasting with corporate platforms whose centralised moderation and opaque algorithms have been shown to cause extensive harms to marginalised communities [38, 47, 104] through biased enforcement, intrusive data extraction, and inconsistent governance [44, 111]. DSM architectures seek to counter these issues by enabling multiple layers of autonomy — open-source code, locally defined governance rules, and instance-level control over data and federation [1, 105, 112]. Consequently, DSM governance is highly localised and heterogeneous, with each instance cultivating its own norms, moderation practices, and privacy expectations.

Prior work shows that administrators and moderators — collectively, "operators" — are often small volunteer teams who carry broad responsibility for sustaining their communities, from enforcing rules and mediating conflicts to shaping norms and maintaining technical infrastructure [59, 108]. This labour is emotionally demanding and safety-critical: operators routinely confront harassment, misinformation, and extreme or illegal material such as child sexual abuse material (CSAM) [95, 100]. Scholarship characterises this work as civic labour, collective sensemaking, and reflective governance rather than mere content removal [21, 70, 91]. These dynamics are especially salient for LGBTQ+ and other marginalised communities, who rely on DSM's autonomy to establish norms often unsupported by mainstream moderation regimes [45, 46, 71, 94].

A defining feature of DSM governance is its federated nature. Harms often originate outside the local server, requiring operators to manage cross-instance risks through blocklists, shared moderation intelligence, and defederation decisions [62, 107]. These processes are largely ad hoc and labour-intensive, shaped by uneven moderation capacity across the network [4]. Operators therefore govern both inward-facing community norms and outward-facing relationships with neighbouring instances—a governance mode that is simultaneously local, distributed, and interdependent.

Decentralisation also extends to data governance. Instance operators control what data is stored, retained, or shared, yet studies show wide variation in privacy policies, consent mechanisms, and expectations of cross-instance data flows [51, 52, 66]. This heterogeneity shapes which forms of automation or infrastructural support are viewed as legitimate and under what conditions communities are willing to adopt them.

Together, these studies illustrate the complex governance landscape in which DSM operators work. What remains underexplored is how they understand or evaluate the potential role of AI within such decentralised, value-sensitive, and federated arrangements.

2.2 AI and Content Moderation on Centralised Platforms

Centralised social media platforms have long relied on automated systems to manage the scale and velocity of user-generated content. AI-based moderation is commonly framed as a scalable solution, promising efficiency and consistency while reducing moderators' exposure to traumatic material [18, 39, 44]. Major platforms deploy machine-learning classifiers, keyword filters, and hashing systems to detect hate speech, harassment, misinformation, spam, and child sexual abuse material (CSAM), making such systems central to contemporary governance infrastructures [2, 35, 68, 90]. These tools are now deeply embedded in enforcement workflows, shaping what platforms define as harmful and how decisions cascade through automated or semi-automated pipelines.

Yet substantial research has shown that automated moderation struggles with linguistic nuance, cultural variation, and contextual interpretation. Hate-speech and toxicity classifiers frequently encode biases from training data, disproportionately flagging African American Vernacular English (AAVE), queer slang, and other minoritised dialects as abusive [28, 47, 87]. Automated systems also misclassify activism, counter-speech, and political expression, especially when dominant groups are being critiqued [47], and particularly in non-English and Global South contexts, where datasets underrepresent local linguistic and cultural norms [31, 97]. These errors reproduce and amplify structural inequalities, contributing to patterns of over-policing and visibility suppression that disproportionately affect marginalised users – including LGBTQ+ communities, Black communities, and sex workers – whose content is more likely to be removed, down-ranked, or "shadowbanned" by opaque algorithms [23, 47, 72, 101]. Such harms have generated deep mistrust toward AI-supported moderation among affected communities [25, 47], shaping broader public perceptions of algorithmic governance. Recent work on commercial content moderation APIs further shows that these services can simultaneously over- and under-moderate group-targeted hate speech across linguistic varieties, raising concerns about whose speech is disproportionately silenced or left unprotected [49].

In response to the limitations of full automation, recent work conceptualises AI moderation as human–AI collaboration rather than replacement. Systems supporting triage, priority queuing, or conditional delegation allocate routine or high-certainty decisions to models while reserving ambiguous or sensitive cases for human moderators [10, 63, 93]. These hybrid workflows emphasise oversight, contestability, and accountability, but they are designed for and evaluated within large centralised platforms where data, policies, and authority are comparatively unified. Taken together, prior research highlights both the capabilities and limitations of AI moderation in centralised environments. How these approaches translate to decentralised settings, where authority is distributed, norms are locally defined, and data boundaries vary across instances, remains relatively underexplored.

2.3 Governing AI: Human-in-the-Loop, Legitimacy, and Data Governance

Research on algorithmic decision-making emphasises that automated systems must remain within human-in-the-loop governance structures, especially in domains requiring contextual interpretation, discretion, or high-stakes judgment. Across fields such as aviation, clinical decision-support, and online platforms, studies consistently document *automation bias*: people often over-trust or defer to automated outputs even when they are incorrect [22, 43]. These findings underpin Human–AI collaboration research arguing that AI should support, not replace, human decision-making, with clear delegation boundaries that reserve irreversible or normatively sensitive actions for people [63]. Work on sociotechnical fairness similarly shows how algorithmic systems reshape accountability, requiring mechanisms for oversight, correction, and contestation [92]. Together, this literature establishes human accountability, reversibility, and constrained autonomy as core principles for legitimate algorithmic governance.

A parallel line of research shows that algorithmic legitimacy depends not only on accuracy but on procedural justice—how decisions are explained, whether affected parties have voice, and whether they can challenge or exit the system. Transparency practices such as logs, rationales, and audit trails shape perceptions of fairness and trust [64]. Studies of content moderation processes further demonstrate that different institutional arrangements — such as contractors, algorithms, expert panels, and participatory mechanisms — are perceived as having varying degrees of legitimacy, with expertise, accountability, and alignment with community values playing a central role [84]. Work on contestability similarly argues that people must be able to understand, dispute, and seek redress for automated decisions, particularly when errors carry social or personal consequences [69]. Empirical studies of moderation appeals and disputes on large platforms show that such mechanisms are often experienced as opaque, constrained, or ineffective in practice, reinforcing perceptions of unilateral platform power [102, 103]. Studies also find that consultation, opt-in participation, and meaningful notice prior to deployment strongly influence perceived legitimacy [96]. Rather than treating moderation as a narrowly technical problem, scholars argue that it should be understood as a broader socio-technical governance practice, encompassing rules, labour, institutional authority, and participation [41]. Related work on online community governance further shows how volunteer moderators develop and enforce locally specific rules and norms within large platforms, resulting in highly heterogeneous governance practices across communities [34]. These insights extend to civic and public-sector AI, where scholars emphasise that transparency alone is insufficient without participatory processes that allow communities to shape system introduction and governance [78]. Prior failures of automated moderation on mainstream platforms (see §2.2) have contributed to cultural scepticism toward AI within many decentralised communities.

A further dimension of AI governance concerns data boundaries, privacy, and consent. Data governance research highlights how AI systems are constrained by principles of purpose limitation and data minimisation, which restrict the reuse of personal data and require processing only what is necessary for a defined task [19, 33, 54, 77, 88]. These principles have gained prominence amid controversies over training AI models on user-generated content, including regulatory actions on X's use of personal data, backlash to LinkedIn's proposed training policies, and wider debates on scraping online posts [27, 81, 82]. Reflecting these tensions, parts

of the Fediverse, including *mastodon.social*, have introduced terms prohibiting the use of user data for AI training [73]. In parallel, work on data sovereignty and community-owned infrastructures argues for locally governed AI systems insulated from external processors, allowing communities to retain meaningful control over data flows and model behaviour [9, 20, 74, 98].

While prior literature establishes broad principles for privacy-preserving and accountable AI, it primarily examines the perspectives of large, centralised organisations or institutional data controllers. Far less is known about how small, volunteer-run, and autonomously governed DSM communities interpret these principles in practice, or how they determine what would constitute legitimate, trustworthy, or acceptable forms of AI involvement, should such systems be considered.

2.4 AI in Decentralised Social Media: Emerging Work and Open Questions

Existing research on decentralised social media (DSM) has examined its distinctive governance dynamics, including local autonomy, volunteer-led moderation, heterogeneous community norms, and cross-instance coordination practices such as blocklisting and defederation [52, 83, 95, 107, 108]. Recent work also highlights infrastructural challenges that arise as independently run instances coordinate governance at scale [53]. However, these studies have not examined how emerging AI technologies are perceived within such arrangements, nor whether communities view them as relevant, desirable, risky, or unnecessary.

Early explorations of automation in DSM contexts have focused on technical feasibility rather than community governance implications. Systems work shows that machine-learning models can be deployed in decentralised settings, for example through federated-learning approaches that avoid centralising user data [109] or rule-based safety tooling adapted to federated infrastructures [61]. Studies of user-facing algorithmic features, such as feed curation, similarly reveal ambivalent receptions reflecting broader cultural scepticism toward automation [67]. While these efforts indicate that automation could be introduced into DSM, they do not examine how operators evaluate such possibilities or the conditions under which automation would be welcomed, constrained, resisted, or rejected.

As a result, several questions remain open. Existing research does not examine whether operators believe AI has a place in DSM governance, how they assess potential benefits and risks, where they draw boundaries around acceptable or unacceptable forms of automation, or what governance mechanisms they would require for any AI system to be viewed as legitimate. More broadly, it is unclear how principles largely developed in centralised organisational contexts, such as human-in-the-loop control, data sovereignty, transparency, or community consent, translate into small, volunteer-run, autonomously governed instances. Addressing these gaps, this study investigates DSM operators' attitudes and concerns toward the use of AI (RQ1), how they understand the areas of need within DSM and whether they see any in which AI could potentially offer support and what forms such support might take (RQ2), and the governance mechanisms they view as necessary to address its associated challenges (RQ3).

3 Methods

The study began with a short preliminary survey, which we used to recruit participants and gather background information. We then conducted 20 one-on-one sessions — framed as semi-structured interviews informed by Research through Design (RtD) and augmented with generative design probes and speculative scenarios — with administrators and moderators of decentralised social media platforms. Moderators in decentralised social media typically focus on content review and enforcing community rules, whereas administrators additionally manage server configuration, federation decisions, and broader governance and technical responsibilities. We use the term "operators" throughout this paper to collectively refer to these two roles. While most participants managed Mastodon servers, the sample also included participants from Pixelfed, PeerTube, Lemmy, Pleroma, and Funkwhale. Participants were based across six countries (USA, UK, Germany, France, the Netherlands, and Costa Rica). These sessions explored decentralised social media operators' attitudes toward AI, the roles they envisioned for it, and the governance boundaries they believed should shape its use in decentralised social media. This study was approved by the Institutional Ethics Review Board at our organisation [name and reference redacted for review].

3.1 Participant Recruitment

We identified potential participants by browsing the Join Mastodon directory[1] and distributed invitations along with a preliminary survey. We reached out to a wide range of general, interest-based, and regional instances, as well as those established to support marginalised or stigmatised groups. In addition, we used snowball sampling, where several participants referred peers from their networks. In the preliminary survey, respondents were asked whether they would be interested in taking part in a one-on-one online session. Those who agreed were contacted by email to arrange a session. Prior to each session, the participant received an information sheet and consent form, which they completed and returned in advance. With consent obtained, all sessions were conducted remotely via videoconference.

A subset of potential participants had previously taken part in an earlier study [citation redacted for review] we conducted on decentralised social media administration. We re-contacted them because their sustained involvement in server-level governance made them especially well positioned to speak to the questions in this study. Many of these returning administrators run instances that explicitly support LGBTQ+ or otherwise marginalised communities. Their perspectives were particularly valuable, as such communities value social media for identity work, community building and support [30, 36, 50, 58, 89], yet often face pervasive hate and targeted abuse [42, 58, 80, 89], with mainstream platforms providing little substantial support in response [99] and instead algorithmically-delivering and -amplifying such harms [24, 58]. Simultaneously LGBTQ+ communities face disproportionate removals, particularly for "grey area" content [47], with negative impacts on desired community building and support [6, 45, 47, 71, 94]. Hence, LGBTQ+ communities value inclusion and self-determination [26], have a history of infrastructuring their own content moderation and safety mechanisms

[1] https://joinmastodon.org/

[57, 85, 99], and have been active adopters of decentralised systems as safer, more autonomous online spaces [108]. Their long-standing governance experience — often shaped by the need to protect vulnerable users [99, 108] — provided valuable insights relevant to the themes explored in this study.

3.2 Preliminary Survey

A total of 27 respondents completed the short preliminary survey, which we used both to assess eligibility and to facilitate preparation for the subsequent sessions. Eligibility required participants to be at least 18 years old and to currently or recently serve as an administrator or moderator of a decentralised social media server. Beyond basic screening, the survey gathered contextual details such as the size and intended communities of participants' instances, their tenure as administrators or moderators. It also included open-ended prompts about the values of their communities, challenges they had encountered in running their instances, the tools they considered useful for addressing those challenges, and their prior experiences with or attitudes toward AI. These responses provided valuable background knowledge and helped shape the focus of the subsequent sessions.

3.3 Main Session

We conducted an exploratory study framed as semi-structured interviews informed by Research through Design (RtD) [37, 86, 110] and augmented with generative design probes and speculative scenarios. We used the topic of content moderation as a situated entry point because prior work [4, 95, 108, 109] has shown that decentralised social media operators experience moderation as a central challenge and have called for more scalable tooling and support, including automated and collaborative approaches. It is a labour-intensive [95], value-sensitive [106, 108], and technically plausible area where AI support is increasingly explored in decentralised servers [109]. In addition, moderation provides operators with concrete experiences and pain points through which they can articulate what AI could meaningfully do, rather than speaking in the abstract about "AI" in general. At the same time, discussions about AI moderation naturally surface broader attitudes, needs, and concerns, allowing deeper implications to emerge organically rather than being imposed a priori. Choosing moderation as a lens therefore grounds our study in real operational challenges while revealing the broader roles and boundaries operators imagine for AI in decentralised social media.

Because the use of AI in decentralised social media remains nascent — with only early experimentation and no widely established expectations or practices around AI-supported moderation or governance [4, 95, 108, 109] — many of the questions we investigate, such as how AI might collaborate with human operators or where appropriate governance boundaries should lie, cannot yet be meaningfully examined through existing deployed systems alone. This aligns with longstanding arguments in Research through Design (RtD) that when technologies are under-defined or situated within under-constrained problem spaces, speculative and generative design materials can serve as productive means for inquiry rather than attempts to evaluate finished systems [37, 110].

To address this challenge, we developed generative design probes used alongside one-on-one semi-structured interviews. Generative probes are well-established in design research as tools that surface implicit assumptions, future-oriented expectations, and value tensions by inviting participants to interpret, critique, and extend deliberately incomplete artefacts [16, 86]. Our probes ranged from near-term feature probes [86], which grounded discussion in plausible present-day mechanisms (e.g., flagging, explanations, customisation), to speculative, design-fiction-inspired scenario probes [29], which depicted possible future AI governance tensions and data-use dilemmas. As shown in prior work, speculative probes can serve as a useful means to prompt critical reflection about emerging sociotechnical systems and to surface governance considerations that participants may not articulate in more traditional interview settings [79]. Consistent with recent work that frames uncertainty around emerging technologies as a generative resource for design and inquiry rather than a methodological limitation [32], these materials enabled operators to articulate imaginaries of AI's potential roles, the boundaries they believed should constrain its behaviour, and the socio-technical mechanisms through which AI might participate in server operations, collaborate with humans, access or be restricted from server data, and be governed within or across decentralised servers.

Each one-on-one session was conducted remotely, and lasted approximately two hours. To support participant comfort, we paused at the midpoint of each session to ask whether the participant would like a 5–10 minute break, and we reminded them that they were welcome to request a break at any time. Each session was facilitated by the lead author, with each participant engaging through an individual digital whiteboard shared only between the facilitator and participant. Sessions were organised into four main components:

3.3.1 Introduction and Warm-up (around 15 minutes). Sessions began with a short introduction, in which the facilitator reiterated the study purpose and invited operators to describe their instances, their roles, and the current challenges of running decentralised social media. This was followed by an open conversation about their existing moderation practices and initial attitudes toward AI. The goal of the warm-up was to establish rapport, ground the discussion in participants' lived experiences, and prepare them for engaging with the generative probes that followed.

3.3.2 Generative Feature Probes (around 45 minutes). To anchor the discussion in plausible present-day mechanisms while still allowing for forward-looking reflection, we introduced a set of near-term generative feature probes — a common RtD technique in which deliberately incomplete artefacts function as prompts for critique, imagination, and extension. These probes drew on prior work and insights from our preliminary survey and included: (1) an automated content-flagging probe reflecting the repetitive and high-volume workload of decentralised social media moderation [4, 95], (2) an explainability probe exploring concerns about opacity and trust in AI-assisted decision-making [55, 76], and (3) a customisation probe investigating how server-level norms and rules might be encoded locally [56, 108].

Each probe was presented visually on a shared digital whiteboard with brief, minimally specified examples. We designed them as provocations rather than solutions: intentionally partial artefacts

intended to reveal assumptions, surface tacit knowledge, and elicit discussion about what would or would not work within operators' diverse contexts. Participants were asked how such features might or might not operate on their instances, what risks or points of failure they anticipated, and how they would adapt or constrain the ideas.

3.3.3 Speculative Scenario Probes (around 45 minutes). To extend the conversation beyond near-term features and probe governance considerations not yet observable in real systems, we introduced two speculative scenario probes. Drawing on design fiction approaches in HCI [11, 29, 65], these short scenarios situated AI within plausible near-future decentralised social media contexts to make potential consequences tangible.

The first scenario depicted an AI that automatically removed a post from a long-standing community member after several minor violations, raising questions about autonomy, accountability, user trust, and legitimacy. The second scenario explored an AI trained on local server data, surfacing tensions around privacy and data governance. We selected the two scenarios to foreground governance dilemmas previously documented in decentralised social media research, as discussed in the Background (§2.1–§2.4 and references therein), and to translate them into the context of AI assistance. We limited the set to two in order to probe these core governance questions in depth within a 45-minute activity.

Each scenario was paired with semi-structured interview questions inviting participants to reflect on issues such as appropriate levels of automation, requirements for explainability or transparency, and the conditions under which AI should or should not be allowed to access or process server data. As with the feature probes, the scenarios were intentionally speculative rather than prescriptive; their purpose was to provoke critical reflection and help participants articulate the governance mechanisms they believe should constrain AI in decentralised social media.

3.3.4 Open-Ended Reflection (around 15 minutes). After engaging with the feature and scenario probes, participants were invited to reflect on any additional considerations relevant to their decentralised social media server. For participants who saw potential value in AI support, this included articulating any capabilities or forms of assistance they would ideally want. For others, particularly those who expressed scepticism or concern about introducing AI into decentralised social media, the discussion instead focused on clarifying their reservations, boundary conditions, and reasons why certain forms of AI might be inappropriate or undesirable.

We placed this open-ended segment at the end of the session so that participants could draw on the reflections developed throughout the earlier components. By this point, they had already considered concrete examples, articulated challenges, and discussed possible risks or governance tensions. This sequencing enabled participants to offer more grounded and considered reflections.

In total, we conducted 20 one-on-one interview sessions with 20 decentralised social media operators. With the exception of one participant (P12), who had three to six months of administrative experience at the time of the survey, all others had been administering or moderating their instances for over a year. Detailed participant information is summarised in Table 1.

P#	Platforms & Roles	Intended Community	Number of Instance Users	Region
P1	Mastodon admin	LGBTQ+	10,000–99,999	North America
P2	Mastodon admin	General	1,000–9,999	Europe
P3	Mastodon admin	General	1,000–9,999	Europe
P4	Mastodon admin	LGBTQ+	1,000–9,999	North America
P5	Mastodon admin	Shared Interest	1,000–9,999	Europe
P6	Mastodon admin	LGBTQ+	100–999	Europe
P7	Mastodon admin	Academics	1,000–9,999	North America
P8	Mastodon admin	LGBTQ+	10–99	North America
P9	Mastodon admin	LGBTQ+	100–999	North America
P10	Mastodon admin & moderator	LGBTQ+	1 ; 1,000–9,999	North America
P11	Mastodon admin	LGBTQ+	10–99	North America
P12	Pleroma admin	Shared Interest	10–99	Europe
P13	Mastodon moderator	General	100,000 or more	Europe
P14	Mastodon admin	Regional	100–999	North America
P15	Pixelfed/PeerTube/Funkwhale admin	Shared Interest	10–99	Europe
P16	Mastodon admin	Regional	10,000–99,999	North America
P17	Mastodon moderator	LGBTQ+	10,000–99,999	Europe
P18	Mastodon moderator	Regional	10,000–99,999	North America
P19	Mastodon/Lemmy admin	General	100,000 or more	Europe
P20	Mastodon admin	LGBTQ+	10,000–99,999	Europe

Table 1: Participants' platforms and roles, intended communities, instance sizes, and regions as reported in the preliminary survey.

We were mindful that administration and moderation work in decentralised social media is often demanding, emotionally taxing, and unpaid. To avoid imposing additional burden on participants, we adopted several measures throughout the study. Sessions were scheduled entirely at participants' convenience, with flexible rescheduling offered without conditions. Each session included an optional midpoint pause and reminders that participants could take breaks or stop the session at any time. Participation was voluntary, and participants were informed that they could withdraw without consequence. Each participant was compensated with a £30 gift voucher to thank them for their time and participation in both the survey and interview components of the study. These measures were designed to respect participants' time and mitigate disruption to their ongoing responsibilities.

3.4 Data Collection and Analysis

We transcribed all interview audio recordings and stored the resulting transcripts in a password-protected folder on the university system. All personally identifiable information was removed during transcription to protect participants' privacy. We conducted a grounded, thematic analysis of the transcripts [8, 12–14]. The coding process began by splitting the transcripts into two equal-sized sets. Two authors independently analysed the first set to develop an initial set of codes. They then met to merge these codes into a shared codebook, discussing and resolving differences through iterative refinement. The first author subsequently applied this codebook to code the remaining set of transcripts.

Rather than aiming for theoretical saturation, we followed recent guidance on interview-based research that emphasises aligning sample size with the depth and scope of the analytic aims [60]. Our goal was to develop an in-depth qualitative understanding of the attitudes, concerns, perceived needs, and governance expectations of decentralised social media operators regarding the use of AI. As this was an exploratory study, the sample we recruited provided a sufficient range of themes relevant to the scope of our work. We

therefore consider the sample size appropriate for the objectives of this research.

4 Results

We structure the results around the study's research questions. Insights related to RQ1 — decentralised social media operators' attitudes and concerns towards AI — run across all subsections. Subsection 4.1 addresses RQ2 by outlining the potential roles and areas of need where AI could support DSM. Subsection 4.2 responds to RQ3, detailing the governance mechanisms participants deemed necessary for AI to be used responsibly and legitimately.

4.1 Potential Roles for AI in Decentralised Social Media

4.1.1 AI as Contextual Intelligence. DSM operators envisioned AI as a source of contextual intelligence, capable of integrating internal community histories with external references to support fairer and more informed decisions. Governance decisions, they stressed, cannot be made without context. Participants' desired support distinguished between *internal context* — the conversational, behavioral, and precedent-based histories within the community — and *external context* — such as factual, political, linguistic, and regulatory references beyond it. For internal context, P1, an LGBTQ+ Mastodon instance admin, explained: "Not only do I want to see the reported message ... but I would like to also see what it was being replied to," noting how queer discourse or reclaimed slurs are easily misread when stripped of their conversational thread. Others imagined more detailed behavioral summaries. P10, who once served as moderator for a large LGBTQ+ Mastodon instance and now run their own single-user Mastodon instance, wished for indicators showing whether "this person has insulted 100 different people or has been fighting with this other user for six months," and P17, an LGBTQ+ Mastodon instance moderator, described this as a "vibe check," a quick read on how a user's long-term behavior aligns with community norms. Several participants extended this notion to moderation logs and precedent, imagining AI summarizing past actions, warnings, or "thin ice" status to ground current decisions in a broader history. Participants also imagined AI acting as a form of collective memory — preserving moderation history that Mastodon normally deletes, as P1 noted when posts "are all going to get deleted" and moderators later "can't remember what exactly they said or did," enabling more consistent decisions over time.

External context was equally critical. P2, a general Mastodon instance admin, highlighted misinformation as a major time sink: when dubious links or conspiracy-like posts appear, moderators currently "go to the website or Wikipedia to cross-check facts," a long and tedious process. They envisioned a fact-checking co-pilot that would pre-analyze posts, flag likely misinformation, and provide a confidence score about the credibility of the source. For participants moderating across cultural or linguistic boundaries, external context also included interpretive cues and translation. As P19, a general Mastodon and Lemmy admin based in Europe, noted that, "If something gets posted that is offensive because of some background in US politics, which I don't know. It would be handy if the AI could tell me what the background is? Why is it offensive?" Several also imagined AI situating posts within wider Fediverse or world events — for instance, interpreting a spike of aggressive posts during a major news incident or identifying when certain problematic groups had recently migrated across servers. Beyond this, P5, a shared-interest Mastodon instance admin, emphasized compliance with external regulations, explaining that their top "knockout criteria" for adopting AI would be any feature that helps ensure legal compliance with the UK Online Safety Act, underlining how admins also imagined AI as a compliance assistant navigating broader regulatory demands. At the same time, participants stressed that contextual support must be *trustworthy and succinct.* They feared AI surfacing irrelevant information or misidentifying which contextual layers actually matter, which would add rather than reduce work.

Together, these accounts illustrate how DSM operators imagined AI as a contextual interpreter, one that weaves together internal and external references, while remaining cautious about whether AI could reliably capture the nuance that human volunteers currently piece together through time-consuming investigation.

4.1.2 AI as a Federation-Level Actor. A distinctive need shared by participants was for AI systems that operate beyond a single community. Participants noted that many moderation issues originate through federation rather than locally, and imagined AI as a network-aware partner that helps instances share intelligence, assess external threats, and navigate inter-instance governance while respecting local autonomy. Several wanted AI to formalise the informal, trust-based information exchanges that currently occur among operators across instances. P1 compared this to email spam filters, where "most email services will learn for everybody," suggesting Mastodon AIs should likewise "learn from different communities" so that harmful patterns discovered elsewhere could be intercepted on first appearance. P5 envisioned "a forum of AI co-pilots" exchanging real-time alerts — "Has anybody got any information on this specific user?" — to which another AI might reply, "Oh yeah, I recognise that guy. We had problem with him." As a Pixelfed, PeerTube, and Funkwhale admin, P15 similarly proposed that when an AI flags harmful cross-instance content, it could draft a report for the local admin to forward to the origin server, maintaining human control while enabling more coordinated response.

Participants also imagined AI as a mechanism for generating federation-level situational awareness. P3, a general-instance admin, imagined metrics showing how often a server is blocked elsewhere or whether its content is dominated by "slurs" or "weird images." P6, an LGBTQ+ instance admin, suggested a dashboard summarizing large-scale signals — "(AI) has scanned 10,000 posts and 1000 of those posts got flagged" — to direct attention to potentially problematic neighbors. P7, an academic instance admin, proposed a "Fediverse health check" aggregating posting surges or waves of defederation. P9, an LGBTQ+ instance admin, observing that "the vast majority of our moderation is moderating content from other instances," envisioned risk scores ranking servers by the prevalence of harmful content. P19, a Mastodon and Lemmy admin, echoed this need for early detection, noting that AI should be able to flag instances where their users produce hateful posts. Together, these scenarios position AI as infrastructure for federated early warning and reputation signals. Furthermore, P5, a UK-based

administrator motivated by compliance with the Online Safety Act and by the aspiration to make the Fediverse safe, argued that "AI is the only way to dramatically improve safety at scale, far beyond the limited resources of volunteer moderators." In their view, the decentralized nature of the network, with "too many moving parts" and "too much reliance on every instance acting in good faith," made automation indispensable: "I don't think there is a way to effectively moderate [the Fediverse] without AI."

At the same time, participants emphasised that federation-level AI would sit within contested political relationships and raised concerns about how such systems might reproduce or amplify existing power imbalances. P15 warned that features resembling "a block list of known offenders or trusted user list" had previously generated "a whole drama," and could be easily evaded or strategically manipulated. Some feared that aggregated AI judgments might recentralize influence; P5 described *mastodon.social* as a "recentralizing force" and argued that visibility into another instance's AI settings would be necessary to contest overly strict classifications — "your AI co-pilot is set to be really strict and that leaves us with free speech concerns." Participants also noted structural limits: some instances lack active moderation or are effectively abandoned, creating vulnerabilities that no AI can fully compensate for. Yet for the most severe harms, especially child sexual exploitation, some participants saw cross-federation AI collaboration as essential despite their broader reservations. Participants' accounts suggest that any federation-level AI should be under conditions of transparency, contestability, and human accountability.

4.1.3 AI for Community Well-Being and Moderator Care. Participants consistently framed AI as a means to sustain both moderator and community well-being by reducing repetitive burden, buffering exposure to extreme harms, and supporting healthier forms of communication and coordination. First, participants highlighted the demoralizing and cognitively draining labor of handling repetitive spam. P3 envisioned an "AI spam manager" that acts when multiple signals appear — "a garbage email, a weird name and numbers, dozens of URLs," enabling the system to "automatically delete spam accounts if like all three or four hooks are set." P4, an LGBTQ+ instance admin, wanted a "magic wand" to detect spam waves so "admins don't all have to see the same stupid spam message again and again," and P6 similarly proposed a fediverse-scale tool to detect "spam campaigns" across the Fediverse. While participants welcomed automation in such routinized areas, they emphasized the need for low false positives and fine-grained control to avoid unnecessary escalation.

On the other hand, participants wanted AI to act immediately on the most serious categories of harm, both to protect users and to limit moderators' emotional exposure. P19 stressed that "if it's anything related to CSAM (Child Sexual Abuse Material) ... it should get detected and deleted right away because we're not online all the time." P16 similarly supported automatic action on clear hate speech, threats, or racial slurs. Well-being also encompassed moderators' exposure to traumatic content, particularly CSAM and violent imagery. P2 described such material as "the most difficult [content] to watch," and P20 noted it can leave "lifelong trauma." Participants therefore wanted AI to blur, quarantine, or summarize harmful material, acting as a protective buffer. In addition, several participants linked rapid response to user well-being: P7 noted that hateful messages such as "you should go kill yourself" can "be potentially distressing to that person" and may "turn them off the Fediverse entirely" if left visible while moderators are offline. Yet participants stressed clear boundaries for automation. Mental-health–related posts, in particular, were widely viewed as requiring human judgment. P10 insisted they should be "flagged" but not auto-deleted because people may be "venting." Many warned that assessing intent or severity in such cases is error-prone and ethically sensitive, making full automation inappropriate. These views delineate a boundary where AI may triage, but not replace, human discretion.

Legal obligations further contributed to stress. P16 explained that compliance with mandatory-reporting laws had long been uncertain: "we finally have a lawyer who has volunteered with us ... but before that we were doing our best, but didn't really know if we were in compliance or not." P19 highlighted that the US required moderators preserve and report CSAM rather than delete it, in contrast to European practices. Such ambiguity was described as an emotional burden rather than a merely procedural one. Participants therefore imagined AI not only detecting CSAM but guiding them through varied reporting requirements.

Sustaining community well-being also required collaborative, low-friction moderation workflows. P2 noted that current Mastodon dashboards "lack sufficient features to help us manage the moderation as a group," prompting calls for shared logs, annotations, and coordinated handoffs. Several participants preferred a single, server-level AI that reflects collective judgment rather than individual assistants. P16 warned that the latter could produce "fragments of decision making," and P17 emphasized that learning from all moderators would reduce inconsistency and "risk of reinforcement bias." These reflections point toward a desire for AI systems that augment collective judgment rather than amplify individual differences.

Communication was another domain of emotional labor where participants sought AI support. P16 described drafting announcements as "an hour and a half of chatting," wanting instead an "AI marketer" to produce accessible, audience-tailored messages. P1 noted that asking users for clarification sometimes triggers defensiveness, whereas if "the computer [AI] asked," the interaction might feel less personal. These visions frame AI not only as an enforcement tool but as an interlocutor that reduces interpersonal friction and preserves community trust.

Finally, participants imagined positive, user-facing features that reframed AI as fostering healthier community life. P12 proposed a "boost bot" that surfaces "thoughtful and important" or "lighthearted and funny" posts; P14 described an opt-in "explore mode" for algorithmic discovery; and P20, emphasizing personalization as care, noted that while their LGBTQ+ instance allows discussion of anti-trans legislation, as a trans user they do not want to see such trauma daily: "If someone talks around that term, doesn't use that term, substitutes a character in a word, the filter can't catch it. But AI can." Some participants also advocated automatic image descriptions for accessibility, while others imagined AI summarizing community notes or providing proportionality visualizations to reduce cognitive load.

Nevertheless, a few participants cautioned that AI itself can threaten community mental health. P11 described cases of "AI psychosis," where "people who are engaged with these chatbots thinking that this statistical chain of words loves them and wants them to awaken the human consciousness," and noted that some systems have encouraged unsafe behavior, including "telling people who have need of psychiatric medication to stop taking their meds." They concluded, "why on earth do we want that in our federate spaces? We don't." These reflections underscore that sustaining well-being involves not only mitigating existing harms but preventing new ones that AI systems may introduce.

4.2 Governance Considerations for AI in Decentralised Social Media

4.2.1 AI as Co-Pilot, Not Autopilot. Across all sessions, participants articulated a foundational governance principle: AI must operate strictly as co-pilots, never as autonomous decision-makers. Participants' accounts reveal a shared governance consideration in which AI advice is welcomed, but AI authority is tightly constrained.

Keep Human in the Loop. Many participants consistently rejected AI actions that bypass human oversight. P10 captured this position directly: "I do think that everything should go through a human, even the most obvious things," invoking an older maxim that "a computer can never like be held responsible. Therefore, a computer must never make a management decision." Similar concerns were echoed by P18, who stated, "I definitely don't like auto delete by policy without anybody looking at it," emphasizing that their community prefers AI to "flag things rather than take action" to avoid harm and unnecessary backtracking. Participants also highlighted the user experience implications of autonomous enforcement. P10 worried that automatic actions would make users "paranoid ... unhappy to be on that server" because they cannot contest or understand automated decisions. Collectively, these perspectives frame human-in-the-loop operation as a necessary governance boundary for AI in decentralised social media.

Limit High-Risk Automation. Many participants drew firm boundaries around automated, high-risk actions — especially those that are irreversible. For example, P20 captured the consensus: "The system should not be able to automatically do any non-reversible action ... the system shouldn't be allowed to delete a post if I can't restore it." For most, actions such as post deletions, user suspensions, or instance blocks were therefore reserved exclusively for human moderators, who remain accountable for such decisions. However, participants also described a narrowly defined exception: extreme illegal content such as child sexual abuse material (CSAM). For these cases, the harm of delayed human review outweighed the risks of automation. P19 explained, "if it's anything related to CSAM material, it will be very handy if it gets detected right away and deleted right away ... sometimes something like that gets reported and we only see it 10 hours later and that is not acceptable." They added that even false positives were tolerable: "an AI tool should take action in these cases, just delete it whether it's false or not." P19 further noted that their Lemmy instance already uses an automated system that "checks images ... and if it thinks it has that content, it deletes it ... we don't see the images anymore." Others accepted the same principle but insisted on reversibility. P20 explained, "there are lots of cases ... where it should be allowed to act without human review. I don't want to evaluate CSAM if it's found it," yet stressed that automatic AI actions must remain reversible. Outside this narrow category, participants reverted to a strict posture: no autonomous bans, no automatic deletions of user posts, and no high-impact enforcement without human review. As P10 put it, "a bot automatically delete" is unacceptable because responsibility becomes untraceable.

Deploy AI in Phases. Many participants described a common deployment pattern: AI must begin with minimal authority and progress only after demonstrating reliability. P6 outlined this approach clearly: "I would ideally initially have it just never act automatically ... so we can figure out if it's trustworthy," further noting that "ideally it would always be manual or at least be temporarily held back." Participants viewed such staging as a governance mechanism rather than a technical precaution. Crucially, AI autonomy — if ever granted — must be earned through a demonstrable track record. P9 stressed, "I want to see a track record that demonstrates that the AI can pick those folks out very accurately," while warning that "the stuff it's learning on its own might not actually ever get quite there, because there's a lot of false positives." This skepticism drove strong preferences for human-guided correction. P18 noted, "I want it to learn and improve ... I definitely want to edit or override the record of wrong explanations so that it can learn." Participants therefore framed AI evaluation as a long-term, human-led governance process in which communities actively shape how AI models interpret norms, apply rules, and develop reliability over time.

Ensure Accountability and Redress. Many participants saw accountability and recovery processes as essential governance mechanisms. Administrators — not AI — must remain responsible for all AI-enabled actions. P1 stated this without ambiguity: "The accountability for the mistake would be the administrator ... I'm not going to pass off my accountability for a moderation mistake to the AI." P11 reaffirmed that "You cannot hold an AI accountable. You cannot hold a snippet of code accountable for its decisions. It is not a human being. The admin is." Because mistakes can have serious consequences — "the user can maybe self harm ... So it can have very big consequences if you make a mistake in judging stuff like that. But still the admin would be responsible." as P19 warned — participants described robust redress expectations: "reverse the deletion ... apologize ... explain what happened," and outline steps to prevent recurrence, such as "we'll probably make the AI less strict after this." (P6) Similarly, P9 emphasized the value of transparent explanations: "I can screenshot that and say here's what happened." Many participants also wanted "rollback mechanisms" for reversibility — e.g., muting instead of deleting — to ensure admins can undo harmful actions.

4.2.2 Community-Centred Customization. Across all sessions, many participants emphasised that AI must be fundamentally community-centred. Rather than adopting general-purpose standards or platform-style norms, participants envisioned AI that faithfully reflect each server's distinctive values, cultures, and rules.

Align AI with Community Norms. Many participants stressed that AI should be aligned with each community's own norms, values, and rules, rather than enforcing generic platform-style standards. Many requested mechanisms for encoding local rules directly

into the system. P1 expressed a common sentiment: "I would like to be able to upload community guidelines for it to follow," adding that if administrators do not upload a document, the AI should "be able to use defaults ... grab what you can publicly find" from the instance's rules. P7 similarly stressed that an AI must "see specifically what's allowed and what's not allowed in our server rules."

Participants highlighted that linguistic and cultural nuances — including queer slang, identity-based expressions, and humor — could be easily misinterpreted by generic models. Several also described real harms, with P9 observing that "AI [is] censoring queer people just because they're queer." P9 explained that "cultural or identity based language ... is context dependent," noting "as a queer person, I can call my friend [a reclaimed slur that would be offensive for heterosexual people to use]." On queer-oriented servers, P10 warned against penalizing identity practices that outsiders misread: "If I'm running a queer server, I don't want to automatically filter reclaimed slurs, because people use those for themselves." P15 similarly stressed that "satire, sarcasm or dark humour is part of life ... we should be able to read if this is a joke or humour." Other communities pointed to local culture and tone, such as P16's observation that "jokes could be seen as more aggressive, but it's just part of the community zeitgeist." Participants also stressed that contextual interpretation is community-specific. What counts as harassment, humour, or playful banter often depends on local histories and relational dynamics — meanings that, as P9 put it, are "context dependent" and must be understood within the community rather than through generic standards.

Some participants also viewed data governance as a site of value expression. Several stressed that AI should learn from community-aligned data rather than from large, generic datasets. P1 argued that models should be "using historical data from our own instance" to correctly interpret local norms, while also drawing selectively on curated external sources such as hate group lists maintained by the Southern Poverty Law Center or shared spam databases. Yet these participants drew firm boundaries around unacceptable training material. As P1 put it, "if [the model] is trained entirely on 4chan then I'll get a grok-like system – something not representative of my community." For them, governing AI meant not only specifying rules but also curating the epistemic inputs the model is allowed to learn from — privileging data that reflects the community's ethos and rejecting sources that could distort or undermine it.

Avoiding imported platform norms was another major concern. As P15 stated, "I do not want the AI to ban everything that the big corporations also ban ... we don't want this," warning that generic models could enforce inappropriate standards. A few participants described equity-oriented approaches that explicitly reject "neutral" moderation. P20 described their model as "intersectional moderation", explaining that "we invert the pyramid" to counterbalance structural inequities, and emphasized that the goal "is to mitigate bias because you assume as true that there is bias in people." This approach highlights how some DSM communities define fairness not as uniformity, but as contextually calibrated enforcement. Together, these accounts underscore that AI must be rooted in the lived norms, values, and culture of individual communities — not external, universalised templates.

Configure and Constrain AI. Given the diversity of norms across the Fediverse, many participants described extensive needs for configurability, interpretability, and administrative control. Many envisioned flexible interfaces that allow tuning sensitivity and enforcement based on content type. P1 requested a "strictness slider per content type ... more lenient for NSFW but less lenient for hate speech." Per-category configuration was considered useful to reflect local boundaries. Many participants also requested keyword-level tools such as banned, allowed, or ignored terms. P6 predicted that "the banned or allowed keywords list is going to be like half the screen because people just find different words to use," while cautioning that "if you leave too many things on the allow list ... spammers are also going to figure that out." P15 similarly emphasized that communities should maintain their own lexical rules rather than inherit platform defaults. Participants also discussed identity-level controls such as block lists or trusted-user lists, though many viewed them as unreliable or potentially unfair, preferring domain-level signals or case-by-case review over fixed identity exemptions. Some administrators, especially those operating multiple services, envisioned exportable community profiles so that finely tuned configurations could be reused across instances or restored after failures.

Being able to interpret and constrain the AI were viewed as core governance requirements. P1 insisted that "if it is taking some sort of action, it should leave a note ... using the reasoning that it did," and P7 stressed corrective feedback: "Being able to tell the AI why that was wrong would be helpful." Many participants expected not only visibility into decisions but the ability to reverse or reshape them, though some thought adaptive learning was acceptable only under explicit, administrator-controlled conditions. P1 described a manual approach where the AI can improve from explicit human feedback, while P3 described automatic learning as a privacy violation, warning that an AI "should not snoop and check past moderations." P15 rejected automated learning outright: "I would not want this. I don't think we are a good data set on moderation ... I would rather [manually] inject things into a database." For these participants, automatic learning was incompatible with their expectations. In line with this caution, P1 emphasized the right to disable misaligned features entirely: "I should be able to turn off the feature if it makes too many mistakes."

4.2.3 Transparency and the Co-Construction of AI Legitimacy. Across the sessions, participants stressed that in decentralised social media communities, AI could gain legitimacy if it was transparent, auditable, and introduced through participatory processes. Transparency was not framed as a technical affordance alone, but as the basis upon which accountability, fairness, and community trust could be established and maintained.

Make AI Explainable and Auditable. Participants consistently rejected opaque or automatic systems in favor of AI whose reasoning could be inspected, checked, and — when necessary — explained to the community. As P7 emphasized, "It's very easy to lose the trust of your users, and it's very hard to regain it," positioning explainability as a core governance requirement rather than a convenience. For many, this meant offering concrete and inspectable decision trails. P14 described the need for transparency tools such as "dashboards, logs, and explanations," signaling the need to surface triggers, contextual cues, and model reasoning.

Many administrators wanted to see precisely which data or heuristics shaped an AI judgment. Some framed this in terms of data sources — "what data is used for that ... their posts, their interactions?" (P6) — while others wanted model-level provenance. P20 argued that any tool "should say this AI is trained to use the following [data such as] all public posts ... all moderation reports," making internal logic legible.

These explanations were not only for moderators' internal use but for public accountability. P7 envisioned posting aggregate statistics — "statistics on so many posts were deleted for hate speech ... so many users were suspended" — as a way to communicate AI-supported decisions to the wider community. Through these mechanisms, participants linked explainability to accountability: visible logs and articulate reasoning ensured that responsibility for AI-assisted actions remained traceable and contestable, rather than displaced onto an opaque system.

Govern AI Adoption with Community Process. Beyond transparency of decisions, participants argued that the process by which AI is introduced must itself be transparent and participatory. For many, legitimacy depended on prior disclosure, opportunities for feedback, and meaningful mechanisms for consent.

Many participants emphasized that introducing AI without announcement or consultation would violate community expectations. P7 noted that "there should be an announcement of that and the opportunity for community feedback," even if not every choice required a vote. Others advocated stronger procedural safeguards. P14 argued that "the community should decide how to embrace some tool like this ... like a community vote and opt-ins," and further suggested offering "a period to allow people to go away or migrate to another server before enabling a tool like that."

Explicit consent surfaced repeatedly. P11 stated that "opt-in is always more trustworthy than opt-out," and that users must have "the opportunity to say nope, I don't consent to that." P15 similarly argued that "you have to ask the users beforehand ... At least tell them in two weeks I will enable AI ... If you don't want to have AI, you can leave the instance," articulating transparency as a form of procedural fairness. P20 framed this as a dual responsibility: while "it would make sense that it's the admins who make these decisions" about configuring the AI, "the community will flip out, so there should be a toggle where users can opt in."

Across sessions, participants emphasized that transparency is a condition for informed choice: in decentralized networks, users decide whether to join, remain in, or leave an instance only when its AI practices are clearly disclosed.

Recognise Fediverse Cultural Resistance to AI. Across the sessions, participants recognised a distinctive cultural backdrop of the Fediverse — one shaped by its long-standing skepticism toward AI. A core source of this skepticism is historical. Participants repeatedly described how members of their communities joined decentralised networks to escape the opaque and heavy-handed algorithmic systems of corporate platforms. As P12 recalled, on mainstream sites "a pretty dumb algorithm just matching some words and saying OK that's a violation," whereas on the Fediverse, "nothing ever gets censored without a human looking at it first." Interactions in the Fediverse is grounded in social ties, shared histories, and contextual understanding within and across instances. Participants viewed AI as potentially disruptive to this relational fabric because automated systems may not be able to completely discern the contextual sensitivity and value alignment that human decision-makers cultivate over time. P11 warned that delegating decisions to automation risks "the erosion of trust ... the loss of human connection ... as we abdicate sound human judgment to dubious algorithmic decision making," a shift that threatens the social cohesion many operators see as the defining strength of decentralised social media.

Memories of past harms from corporate AI systems further amplify this skepticism. Participants recalled examples in which automated moderation misinterpreted context or disproportionately silenced marginalised voices. As P9 noted in describing misclassification harms, "AI [is] censoring queer people just because they're queer." P12 offered the example of COVID-related posts that were "getting stripped by algorithms," reinforcing a collective sense that automated systems routinely misjudge nuance. These shaped a cultural narrative within the Fediverse in which AI is associated with misclassification, overreach, and inequity.

Some participants also expressed concern that AI may threaten the authenticity and meaning of community spaces. P11 envisioned worst-case scenarios in which generative systems flood timelines with synthetic content: "Federated timelines being so full of slop that nobody can trust they are reading human thoughts," producing a world where "more machine-generated content than user-generated content and so the users just leave." For these operators, AI does not merely automate — it potentially destabilises the human-centered culture that defines the Fediverse.

Several participants framed resistance to AI within a broader critique of the societal narrative that AI is inevitable. P14 argued that "we are using AI out of fear of being left behind ... big companies are selling us the idea that if you don't start using some tools you will become obsolete," noting that such decisions are "not completely rational as they are based on fear." Others expressed discomfort with extractive or plagiaristic uses of data, as P11 noted when explaining that "people are not going to go out of their way to give consent for their words to be plagiarised." These reflections underscore that cultural resistance is also a resistance to dominant technological imaginaries.

Together, these accounts depict Fediverse skepticism toward AI as a deeply embedded cultural condition rather than a temporary hesitation or isolated preference. This resistance is shaped by the network's origins, reinforced by relational governance norms, and intensified by memories of algorithmic harm and broader critiques of AI inevitability. Participants thus framed legitimacy not in terms of accuracy alone, but in terms of alignment with the human-centred, context-sensitive and autonomy-protecting values they associated with decentralised social media.

4.2.4 Privacy and Data Governance Constraints. Participants consistently framed privacy and data governance as constraints that define the legitimate scope of AI in decentralised social media. These constraints operate at multiple layers, and collectively determine what AI can access, where it should run, how its outputs must be limited, and under what conditions it may operate without violating the social contract of the Fediverse.

Keep Data and Models Local. Many participants expressed that AI should operate under strict local control, with data remaining

on the local server. They described sending content to cloud APIs as a fundamental violation of the decentralised ethos. As P2 put it, "The AI has to be hosted by us and managed by us so that no content can go out of the server." Others echoed that trust in Mastodon rested on precisely this guarantee: "Make sure that the data isn't going outside ... it's local only" (P8). Several warned that routing private data to external services would be intolerable: "You can immediately close your instance if anyone knows about this" (P15). Locality was not only a privacy measure but a cultural and political stance — community data belonged to the community. Nevertheless, while many demanded strict local self-hosting, some acknowledged pragmatic trade-offs. A few participants saw value in pooling curated external datasets, but only when they aligned with community norms. Some thought that cloud services could be acceptable under rigorous safeguards, and should have encryption and plain-language disclosure. A few envisioned hybrid futures in which some communities might eventually run cooperative AI infrastructure.

Set Data Boundaries. Participants' accounts drew multiple, sometimes contested, boundaries around what kinds of content AI could legitimately access. Certain categories such as deleted content, private messages, and follower-only posts were widely regarded as inviolable. P6 was categorical: "No it should never train on that because deleted and private are very specific words that tell you something is not to be shared or it shouldn't exist anymore." P7 similarly refused AI access to private messages: "Direct messages, no. Personally and ethically, I don't think we should do that," while P19 added that AI should only see DM content if it was explicitly reported, mirroring the same expectations for human moderation.

Participants diverged on whether public posts constituted legitimate training data. Some adopted a pragmatic stance as P4 put it, "It's on the Internet. It's there forever. If you post it publicly, you can't get mad when somebody uses it." P18 and P19 echoed that public timelines could be available for AI training, since anyone could already scrape them. Others, however, rejected this view. P9 recalled community outrage when researchers scraped public Mastodon posts without consent: "The moment you take actions like that without the consent of your user base, you lose your user base." P16 also pointed out that "public" did not imply "consent": cross-instance users had not agreed to have their posts used for another server's AI, and using federated content for training raised distinct jurisdictional boundaries.

A further tension concerned how different kinds of "removed content" should be treated. Many participants insisted that content deleted by users must never persist in training data. Yet several felt that content removed by moderators — especially for rule violations — could be valuable training data. P7 explained, "Admin deleted posts, I think I probably would want [the AI] to do that, but we should be transparent about it." P16 similarly drew a crisp distinction: "If it's us taking action [to remove content] that's data that it should train on. If it's the user taking action to remove content I would say no."

Participants also differed on whether moderation logs and user reports were appropriate data sources. Some saw them as the most relevant and least contentious training inputs. Others viewed them as deeply sensitive: P9 argued that exposing report data — even internally through an AI — would be unacceptable, warning that models might later repeat or leak sensitive details. These reflected broader concerns over the confidentiality of moderation work and the risks of embedding sensitive information into AI.

Some participants emphasised that users themselves should have control over whether their content enters training datasets, advocating user-level opt-ins or opt-outs that allow individuals to determine whether and how their data may be used, thereby extending data boundaries to the level of individual agency.

Control Data Sharing. Participants emphasised controlling the circulation of data — both within and across servers. Many wanted granular admin controls such as a checklist or toggle system allowing them to specify which data sources the AI could use. Others proposed that rather than sharing raw posts, instances could exchange aggregated signals like spam hashes or hate keywords. When discussing potentially beneficial cross-instance intelligence sharing, a few participants resisted mechanisms that would expose raw user content. P9 recounted past incidents where instances allowed bots "scraping all public posts and archiving them somewhere," noting that "a lot of people see that as a violation of their privacy" and that their community would "suspend instances that do that." Instead of raw text, participants preferred controlled, minimal, or aggregated signals. P5 described a model where the AI processes private attachments locally but "only passes on information to the moderators if it finds a problem," ensuring that content itself does not circulate unnecessarily. Others warned that even internally-stored materials such as mod reports or admin logs should not be reused without explicit justification. For example, P15 stressed that local data storage "doesn't mean it's yours to use," underscoring that constraints are tied to normative expectations about purpose limitation rather than mere data location.

Respect User Agency. Participants argued that decisions about data use required transparency and legitimate community governance. For some, this meant administrative discretion informed by user expectations; for others, it required explicit community voting or consultation. Regardless of the mechanism, participants described training an AI on local data without clear disclosure as a breach of privacy. P11 insisted that "opt in is always more trustworthy than opt out ... If the ethical thing to do is going to sabotage your tech, then maybe you should consider a different tech." Similarly, P13 insisted: "If the person is against [it], then you don't use those [posts]". P15 similarly stressed that major changes should be publicly announced in advance, with time for users to object, leave, or request deletion. Transparency features such as data policies, training disclosures, and dashboards were recommended to reassure skeptical communities. A few participants also highlighted the importance of mechanisms that ensure deleted content does not persist indefinitely in AI systems. Across communities, consent, visibility, and process integrity were treated as integral to legitimate data governance. Together, these expectations define data governance constraints that AI systems must operate within to remain legitimate in decentralised social media.

5 Discussion

5.1 Perceived Legitimacy of AI in Decentralised Social Media

Our findings show that decentralised social media (DSM) operators evaluated AI through the lens of sociotechnical legitimacy — whether its operation aligns with DSM's governance norms, cultural expectations, and relational practices. Legitimacy rested on several conditions. First, accountability had to remain visible, human, and instance-specific: regardless of whether AI supported content handling, coordination, or cross-server insights, responsibility had to trace back to administrators rather than opaque automation. Reversibility, contestability, and explainability were core expectations, echoing long-standing findings that automated systems should augment rather than replace contextual human judgment [22, 43, 63]. Procedural governance was equally important. Operators expected advance disclosure, opportunities for feedback, and meaningful exit options — reflecting the distributed agency of DSM users, whose ability to choose and leave instances is foundational to the ecosystem. These expectations align with procedural justice frameworks emphasising transparency, voice, and contestability [64, 69, 78, 96]. Finally, AI had to respect instance-defined norms. Because DSM is governed by plural, community-specific cultures rather than a single rulebook, uniform AI behaviour risked undermining the autonomy decentralisation was designed to protect.

These expectations are rooted in the Fediverse's cultural history [1, 53, 108]. Participants contrasted DSM with corporate platforms, framing decentralisation as a move away from opaque, unilateral, and automated governance [40, 111]. Many invoked past algorithmic harms — misclassification of marginalised groups, unaccountable enforcement, intrusive data practices, and generative spam — to explain their heightened scepticism. This sensitivity to information flow echoes analyses of "trust and friction" in DSM [52] and reflects well-documented moderation harms, including misclassification of queer language and disproportionate suppression of marginalised users [23, 28, 47, 87, 101]. Work on decentralised feed curation similarly shows resistance to opaque algorithms [67], and studies highlight the limits of formal transparency mechanisms [17]. These histories contribute to the broader cultural resistance shaping how DSM communities approach AI.

In DSM, AI must be legitimate before it can be useful. Legitimacy delineates what forms of AI are acceptable or even imaginable, shifting adoption from a technical problem to one of governance — aligning AI with community autonomy, procedural fairness, and the cultural commitments of the Fediverse. This framing underpins operators' imagined roles for AI and the governance mechanisms discussed in the sections that follow.

5.2 Reimagining AI in DSM: Contextual, Federated, and Care-Oriented Governance Infrastructure

Our findings reveal a distinctive imaginary of AI's role in decentralised social media (DSM). Rather than viewing AI as a classifier or autonomous decision-maker developed for centralised platforms [18, 39, 44, 90], operators conceptualised it as governance infrastructure — augmenting human oversight, reinforcing community autonomy, and supporting the relational practices that sustain the network. This imaginary departs from platform-centric paradigms and emerges from DSM's structural and cultural specificities; even research on hybrid or collaborative moderation [10, 93] presumes organisational unity that DSM lacks.

Operators envisioned AI as contextual intelligence: an infrastructural system that assembles and surfaces the histories, relationships, behavioural patterns, and external references that underpin governance. Existing tools present posts in isolation, detached from conversational threads, norms, and precedents. AI was therefore imagined as a means to restore context — reconstructing discussions, recalling past decisions, interpreting culturally specific expressions, summarising behavioural histories, and situating events in broader social, legal, or linguistic frames. This role aligns with guidelines emphasising contextual explanations and decision support [3, 55]. In DSM, where information is fragmented across servers and lacks unified visibility layers, operators must actively reconstruct context; AI thus becomes a facilitator of informed, situated judgement.

Operators also articulated a federated role for AI grounded in the relational nature of the Fediverse. Governance challenges often arise across instances rather than within them — through content flows, neighbouring servers' behaviours, shifting reputations, and waves of new users or external events. AI was envisioned to support this layer by surfacing cross-instance signals, identifying emerging risks, mapping defederation patterns, and providing high-level network "health indicators." This reflects the difficulty of constructing cross-instance visibility in a fragmented ecosystem, as seen in infrastructural work like FediLive [75], and aligns with analyses of the political choices encoded in DSM protocols [83] and research on federated governance [4, 53, 62, 107]. Crucially, such tools must avoid becoming mechanisms of recentralisation; AI should support, not homogenise, instance autonomy.

Finally, operators imagined AI as a tool for community and operator well-being. They emphasised the emotional and relational labour of maintaining community spaces — exposure to harmful content, burnout from repetitive tasks, drafting sensitive communications, coordinating distributed teams, and supporting vulnerable users. This builds on work conceptualising moderation as emotional and civic labour [70, 91, 95], extending it by envisioning AI as a buffer or collaborator: reducing exposure to distressing material, detecting spam or conflict, assisting communication, and enabling personalised filtering. Yet they also cautioned that AI could undermine well-being through generative misinformation, trust erosion, or harmful conversational agents, underscoring that care must extend both to and around AI.

Taken together, these visions position AI not as an automated decision-maker but as a distributed governance partner — providing contextual intelligence, enabling federation-level coordination, and contributing to collective care. This imaginary aligns with DSM's architecture of local autonomy and cross-instance interdependence and challenges AI governance models that assume centralised control, uniform norms, and global data access. The next section examines the governance mechanisms operators view as necessary to sustain such forms of AI within decentralised environments.

5.3 Governing AI in DSM: Negotiating Autonomy, Transparency, and Data Boundaries

While operators described expansive visions for how AI might support decentralised social media (DSM), they also articulated the governance boundaries under which such systems could be responsibly integrated. These were not uniform rules but negotiated boundaries shaped by instance autonomy, community norms, and the structural properties of federation.

DSM operators insisted that AI remain under human and instance-level control, reflecting the ecosystem's foundational commitment to autonomy. AI was consistently framed as a co-pilot whose authority derives from, and is constrained by, administrators. Participants emphasised visible responsibility, reversible actions, and gradual, supervised deployment that allows AI to "earn" trust, paralleling findings on appeals and incremental integration [5, 15, 55]. These expectations extend beyond specific tasks: they articulate a broader principle that AI must remain subordinate to human judgement and local governance. This stance aligns with HCI principles emphasising human primacy and reversibility [3, 7], and echoes decentralisation's ethos of resisting top-down algorithmic authority.

Operators also stressed the need for community-centred configuration, reflecting the heterogeneity that defines the Fediverse. Instances vary widely in cultural norms, linguistic practices, political orientations, and expectations of care; configuration was therefore framed as a governance right rather than a technical feature. Operators expected systems to incorporate instance-specific rules, support fine-grained tuning, adapt to contextual practices such as reclaimed slurs, and allow communities to disable misaligned features. Normative diversity is structural in DSM, and governance mechanisms must equip instances, not platforms, to define how AI interprets their values. These expectations were especially salient for LGBTQ+ communities, who have long experienced automated moderation that misrecognises their norms [25, 46, 57, 85, 99]. While LGBTQ+-oriented instance operators stressed community-specific configuration, many LGBTQ+ users' deep mistrust of automated systems [47, 48, 58] highlights an important area for future work.

Operators also identified data access, training practices, and learning processes as contested governance domains rather than technical details. They strongly preferred keeping data and models local, extending principles of data sovereignty and community-controlled infrastructures [9, 20, 74] to AI training. Deleted posts, private messages, and sensitive moderation records were widely considered off-limits. These concerns have gained visibility as some Fediverse communities formally prohibit using their content for model training [73], echoing data-minimisation and purpose-limitation principles in regulatory analyses [19, 33, 54]. Other data types, such as public posts or removed content, were subject to negotiation, reflecting contested boundaries documented in decentralised infrastructures [51, 52]. AI learning itself was similarly fraught: some supported limited, observable adaptation grounded in explicit feedback, while others entirely rejected automatic learning due to concerns about privacy, shifting behavioural thresholds, or erosion of community agency. Across these positions, a shared principle emerged: AI learning must be authorised, constrained, and explainable, operating within each instance's norms and consent structures.

Together, these negotiations show that AI in DSM must be governed across multiple layers: control, ensuring accountability to human decision-makers; configuration, enabling communities to encode their values; and consent and data governance, defining what information AI may use and how its behaviour may evolve. These mechanisms reflect the underlying architecture of decentralisation, where communities hold both the autonomy and the responsibility to shape the systems mediating their social spaces. The next section outlines design implications for AI compatible with decentralised, federated environments.

5.4 Design Implications for AI in Decentralised Social Media

While early explorations demonstrate that automation is feasible in DSM [61, 67, 109], our findings reveal the governance conditions under which such systems would be viewed as legitimate. Designing AI for decentralised social media therefore requires rethinking assumptions inherited from centralised platforms. Rather than exporting models optimised for scale, uniformity, or centralised control, systems must honour the autonomy, heterogeneity, and federated interdependence that define DSM. We outline three implications for design.

First, AI should be designed as instance-governed rather than platform-governed technology. In DSM, each community operates as an autonomous sociotechnical unit with its own values and interpretive conventions. AI must therefore support per-instance rule ingestion, instance-specific thresholds, and representations of cultural and social nuance that centralised models routinely flatten. Systems should allow administrators to disable or override features, adjust behavioural parameters, and align outputs with local norms, even when these norms diverge widely. Designing for instance governance challenges assumptions that AI should embody universal standards; instead, AI must enable communities to enact their own governance choices.

Second, AI should support federated governance by increasing visibility and coordination across instances — without recentralisation. Operators envisioned AI as infrastructure that helps communities navigate inter-instance relationships, offering early-warning signals, network health indicators, and privacy-preserving pattern sharing to anticipate emerging risks. These capabilities must be balanced with safeguards that prevent large or well-resourced instances from becoming de facto central authorities. Designing for federated governance therefore requires systems that facilitate awareness and cooperation without reproducing platform-like power asymmetries or compromising local autonomy.

Third, AI learning and data use must be treated as community-governed processes. Decisions about what data AI may access, how models may adapt, and when learning is appropriate are governance questions in DSM. Systems should provide mechanisms for administrators to authorise or restrict data sources; determine whether, how, and on what basis models may learn; and communicate these decisions transparently. Learning should be observable, reversible, and grounded in explicit human feedback, respecting

community-defined boundaries around logs, reports, deleted content, and other sensitive information. Treating learning as a governed process recognises the negotiated nature of data autonomy in DSM and ensures that model evolution does not silently reshape community norms.

Together, these implications imagine AI as decentralised governance infrastructure — configurable by each instance, responsive to federated dynamics, and accountable to the communities it serves. Designing for DSM thus requires aligning AI with the principles that underpin decentralisation itself: autonomy, heterogeneity, transparency, and community control over the systems that mediate their shared social spaces.

6 Limitations

This study has several limitations. For practical purposes, we had to limit our study sample to twenty participants. Rather than aiming for representativeness or thematic saturation, our goal was to get an initial set of exploratory insights which we could examine in depth. While most participants were from Mastodon, reflecting its prominence in the Fediverse, their perspectives may not fully represent experiences on other decentralised platforms with different cultures and affordances. Another limitation stemmed from the fact that all participants were administrators or moderators, often with higher technical literacy and stronger commitments than ordinary users, and all were based in Europe and North America, which may not capture perspectives from other regions and linguistic contexts. In scope, we examined the attitudes, design needs, and governance considerations, but not technical feasibility, implementation barriers, or end-user acceptance. Gender demographics were not collected, as this information was outside the scope of our research questions and analytic focus. Methodologically, our workshops elicited perspectives and reflections, but not observed practices. Our focus on individual administrators across diverse instances provided cross-community breadth, but did not include a community-level investigation, which future work could pursue to capture internal dynamics more deeply. Despite efforts to reach a wide range of instances, relatively few from racially minoritised backgrounds participated in our study. We therefore treat our findings as reflecting the perspectives of those who chose to participate, while acknowledging that further engagement with racially diverse communities remains an important area for future work. Many participants operated LGBTQ+ or marginalised-community instances, reflecting their substantial presence among active DSM operators and the particular governance burdens such communities face. This emphasis may shape the perspectives represented and may limit the generalisability of certain insights.

7 Conclusion

This paper offers the first in-depth account of how decentralised social media (DSM) operators imagine the place of artificial intelligence. Across 20 interviews, operators rejected AI as an autonomous authority and instead envisioned it as governance infrastructure that provides contextual intelligence, supports cross-instance coordination, and helps sustain community and moderator well-being. They also articulated strict boundaries rooted in DSM values: human accountability, reversibility, transparency, community-centred configuration, and strong data-governance constraints. These findings challenge assumptions from centralised platforms and show that responsible AI in DSM requires systems that are instance-governed, configurable, and interpretable, and that support federated coordination without recentralising power.

References

[1] Roscam Abbing et al. 2023. Decentralised social media. *Internet Policy Review* 12, 1 (2023).

[2] Ahlam Alrehili. 2019. Automatic hate speech detection on social media: A brief survey. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 1–6.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[4] Ishaku Hassan Anaobi, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola, and Gareth Tyson. 2023. Will Admins Cope? Decentralized Moderation in the Fediverse. In *Proceedings of the ACM Web Conference 2023*. 3109–3120.

[5] Shubham Atreja, Jane Im, Paul Resnick, and Libby Hemphill. 2024. AppealMod: Inducing Friction to Reduce Moderator Workload of Handling User Appeals. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–35.

[6] Laima Augustaitis, Leland A. Merrill, Kristi E Gamarel, and Oliver L. Haimson. 2021. Online Transgender Health Information Seeking: Facilitators, Barriers, and Future Directions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 205, 14 pages. doi:10.1145/3411764.3445091

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 2429–2437.

[8] Kathy Baxter, Catherine Courage, and Kelly Caine. 2015. *Understanding your users: a practical guide to user research methods*. Morgan Kaufmann.

[9] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021).

[10] Lindsay Blackwell. 2025. Content moderation futures. *arXiv preprint arXiv:2509.09076* (2025).

[11] Julian Bleecker. 2009. Design Fiction: A Short Essay on Design. *Science, Fact and Fiction* 49 (2009).

[12] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. sage.

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[14] Virginia Braun and Victoria Clarke. 2021. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and psychotherapy research* 21, 1 (2021), 37–47.

[15] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. 2023. Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–21.

[16] Sena Çerçi, Marta E. Cecchinato, and John Vines. 2021. How design researchers interpret probes: Understanding the critical intentions of a designerly approach to research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[17] Tyler Chang, Joseph J Trybala III, Sharon Bassan, and Afsaneh Razi. 2025. Opaque Transparency: Gaps and Discrepancies in the Report of Social Media Harms. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[18] Nafia Chowdhury and Daphne Keller. 2022. *Automated Content Moderation: A Primer*. Technical Report. Program on Platform Regulation, Stanford Cyber Policy Center. White paper.

[19] CNIL. 2025. *AI and GDPR: CNIL publishes new recommendations to support responsible innovation*. https://www.cnil.fr/en/ai-and-gdpr-cnil-publishes-new-recommendations-support-responsible-innovation Accessed 2025-12-03.

[20] Nick Couldry and Ulises A Mejias. 2019. The costs of connection: How data is colonizing human life and appropriating it for capitalism. In *The costs of connection*. Stanford University Press.

[21] Amanda LL Cullen and Sanjay R Kairam. 2022. Practicing moderation: Community moderation as reflective practice. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW1 (2022), 1–32.

[22] Mary L Cummings. 2017. Automation bias in intelligent time critical decision support systems. In *Decision making in aviation*. Routledge, 289–294.

[23] Daniel Delmonaco, Samuel Mayworm, Hibby Thach, Josh Guberman, Aurelia Augusta, and Oliver L Haimson. 2024. " What are you doing, TikTok?": How

Marginalized Social Media Users Perceive, Theorize, and" Prove" Shadowbanning. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–39.

[24] Michael Ann DeVito. 2022. How Transfeminine TikTok Creators Navigate the Algorithmic Trap of Visibility Via Folk Theorization. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 380 (nov 2022), 31 pages. doi:10.1145/3555105

[25] Michael Ann DeVito, Ashley Marie Walker, and Julia R Fernandez. 2021. Values (mis) alignment: Exploring tensions between platform and LGBTQ+ community design values. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[26] Michael Ann DeVito, Ashley Marie Walker, and Julia R. Fernandez. 2021. Values (Mis)Alignment: Exploring Tensions Between Platform and LGBTQ+ Community Design Values. 5, CSCW1, Article 88 (apr 2021), 27 pages. doi:10.1145/3449162

[27] Digital Watch – Geneva Internet Platform. 2024. *Open Rights Group slams LinkedIn for data use in AI without consent.* https://dig.watch/updates/open-rights-group-slams-linkedin-for-data-use-in-ai-without-consent Accessed 2025-12-03.

[28] Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful speech detection by language models exhibits gender-queer dialect bias. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.* 1–12.

[29] Anthony Dunne and Fiona Raby. 2024. *Speculative Everything, With a new preface by the authors: Design, Fiction, and Social Dreaming.* MIT press.

[30] Brianna Dym, Jed R. Brubaker, Casey Fiesler, and Bryan Semaan. 2019. "Coming Out Okay": Community Narratives for LGBTQ Identity Recovery Work. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 154 (Nov. 2019), 28 pages. doi:10.1145/3359256

[31] Mona Elswah, Aliya Bhatia, and Dhanaraj Thakur. 2025. Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages. https://cdt.org/insights/content-moderation-in-the-global-south-a-comparative-study-offour-low-resource-languages/

[32] Felix Anand Epp, Anton Poikolainen Rosén, Antti Salovaara, and Camilo Sanchez. 2024. Uncertainties as Generative Resources in Research through Design: Three Dynamics for Moving in a Design Space. *ACM Transactions on Computer-Human Interaction* 31, 6 (2024), 1–31.

[33] European Data Protection Board. 2024. *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models.* Technical Report. European Data Protection Board. Accessed 2025-XX-XX.

[34] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[35] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)* 51, 4 (2018), 1–30.

[36] Jesse Fox and Rachel Ralston. 2016. Queer identity online. *Comput. Hum. Behav.* 65, C (Dec. 2016), 635–642. doi:10.1016/j.chb.2016.06.009

[37] William Gaver. 2012. What should we expect from research through design?. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 937–946.

[38] Christine Geeng and Alexis Hiniker. 2021. LGBTQ privacy concerns on social media. *arXiv preprint arXiv:2112.00107* (2021).

[39] Tarleton Gillespie. 2020. Content Moderation, AI, and the Question of Scale. 7, 2 (2020). doi:10/gjt335

[40] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[41] Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T Roberts, Aram Sinnreich, and Sarah Myers West. 2020. Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review* 9, 4 (2020), 1–29.

[42] GLAAD. 2024. *Unsafe: Meta Fails to Moderate Extreme Anti-trans Hate Across Facebook, Instagram, and Threads.* Technical Report. GLAAD. https://glaad.org/smsi/report-meta-fails-to-moderate-extreme-anti-trans-hate-across-facebook-instagram-and-threads/

[43] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.

[44] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[45] Rachel Griffin. 2024. The Heteronormative Male Gaze: Experiences of Sexual Content Moderation Among Queer Instagram Users in Berlin. *International Journal of Communication* 18, 0 (2024), 1266—-1288. https://ijoc.org/index.php/ijoc/article/view/21576

[46] Oliver L Haimson, Justin Buss, Zu Weinger, Denny L Starks, Dykee Gorrell, and Briar Sweetbriar Baron. 2020. Trans time: Safety, privacy, and content warnings on a transgender-specific social media site. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[47] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[48] Oliver L. Haimson, Samuel Reiji Mayworm, Alexis Shore Ingber, and Nazanin Andalibi. 2025. AI Attitudes Among Marginalized Populations in the U.S.: Nonbinary, Transgender, and Disabled Individuals Report More Negative AI Attitudes. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25).* Association for Computing Machinery, New York, NY, USA, 1224–1237. doi:10.1145/3715275.3732081

[49] David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in moderation: How commercial content moderation apis over-and under-moderate group-targeted hate speech and linguistic variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.* 1–26.

[50] Lynne Hillier, Kimberly J. Mitchell, and Michele L. Ybarra. 2012. The Internet As a Safety Net: Findings From a Series of Online Focus Groups With LGB and Non-LGB Young People in the United States. *Journal of LGBT Youth* 9, 3 (2012), 225–246. doi:10.1080/19361653.2012.684642

[51] Sohyeon Hwang, Priyanka Nanayakkara, and Yan Shvartzshnaider. 2023. Whose Policy? Privacy Challenges of Decentralized Platforms. In *CHI'23 Workshops: Designing Technology and Policy Simultaneously: Towards A Research Agenda and New Practice.*

[52] Sohyeon Hwang, Priyanka Nanayakkara, and Yan Shvartzshnaider. 2025. Trust and Friction: Negotiating How Information Flows Through Decentralized Social Media. *arXiv preprint arXiv:2503.02150* (2025).

[53] Sohyeon Hwang, Sophie Rollins, Thatiany Andrade Nunes, Yuhan Liu, Richmond Wong, Aaron Shaw, and Andrés Monroy-Hernández. 2025. Governing Together: Toward Infrastructure for Community-Run Social Media. *arXiv preprint arXiv:2509.19653* (2025).

[54] Information Commissioner's Office. 2025. *How should we assess security and data minimisation in AI?* https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/ Accessed 2025-12-03.

[55] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[56] Shagun Jhaver, Seth Frey, and Amy X Zhang. 2023. Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms. *Social Media+ Society* 9, 4 (2023), 20563051231207857.

[57] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (mar 2018), 33 pages. doi:10.1145/3185593

[58] Kay Kender and Katta Spiel. 2025. Social Media as Marginalisation Machine: The Trans Desire for Solidarity Spaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 955, 17 pages. doi:10.1145/3706598.3713136

[59] Erin Kissane and Darius Kazemi. 2024. Findings report: Governance on Fediverse microblogging servers.

[60] Eleanor Knott, Aliya Hamid Rao, Kate Summers, and Chana Teeger. 2022. Interviews in the social sciences. *Nature Reviews Methods Primers* 2, 1 (2022), 73.

[61] Lucio La Cava and Andrea Tagarelli. 2025. Safeguarding decentralized social media: Llm agents for automating community rule compliance. *Online Social Networks and Media* 48 (2025), 100319.

[62] Samantha Lai, Yoel Roth, Renée DiResta, Kate Klonick, Mallory Knodel, Evan Prodromou, and Aaron Rodericks. 2025. *New Paradigms in Trust and Safety: Navigating Defederation on Decentralized Social Media Platforms.* Technical Report. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2025/03/fediverse-social-media-internet-defederation Published March 25, 2025.

[63] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–18.

[64] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–26.

[65] Joseph Lindley and Paul Coulton. 2015. Back to the future: 10 years of design fiction. In *Proceedings of the 2015 British HCI conference.* 210–211.

[66] Mareike Lisker and Helena Mihaljević. 2025. Data Ethics in the Fediverse: Analyzing the Role of Instance Policies in Mastodon Research. *arXiv preprint arXiv:2505.07606* (2025).

[67] Yuhan Liu, Emmy Song, Owen Xingjian Zhang, Jewel Merriman, Lei Zhang, and Andrés Monroy-Hernández. 2025. Understanding Decentralized Social Feed Curation on Mastodon. *arXiv preprint arXiv:2504.18817* (2025).

[68] Emma Llansó, Joris Van Hoboken, Paddy Leerssen, and Jaron Harambam. 2020. Artificial intelligence, content moderation, and freedom of expression. (2020).

[69] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[70] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.

[71] Samuel Mayworm, Kendra Albert, and Oliver L. Haimson. 2024. Misgendered During Moderation: How Transgender Bodies Make Visible Cisnormative Content Moderation Policies and Enforcement in a Meta Oversight Board Case. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 301–312. doi:10.1145/3630106.3658907

[72] Dawn McAra-Hunter. 2024. How AI hype impacts the LGBTQ+ community. *AI and Ethics* 4, 3 (2024), 771–790.

[73] Ivan Mehta. 2025. *Mastodon updates its terms to prohibit AI model training.* https://techcrunch.com/2025/06/17/mastodon-updates-its-terms-to-prohibit-ai-model-training/ Accessed 2025-12-03.

[74] Stefania Milan and Emiliano Treré. 2019. Big data from the South (s): Beyond data universalism. *Television & New Media* 20, 4 (2019), 319–335.

[75] Shaojie Min, Shaobin Wang, Yaxiao Luo, Min Gao, Qingyuan Gong, Yu Xiao, and Yang Chen. 2025. FediLive: A Framework for Collecting and Preprocessing Snapshots of Decentralized Online Social Networks. In *Companion Proceedings of the ACM on Web Conference 2025*. 765–768.

[76] Maria D Molina and S Shyam Sundar. 2022. When AI Moderates Online Content: Effects of Human Collaboration and Interactive Transparency on User Trust. 27, 4 (2022), zmac010. doi:10.1093/jcmc/zmac010

[77] Rainer Mühlhoff and Hannah Ruschemeier. 2025. Updating purpose limitation for AI: a normative approach from law and philosophy. *International Journal of Law and Information Technology* 33 (2025), eaaf003.

[78] Dave Murray-Rust, Kars Alfrink, and Cristina Zaga. 2025. Towards Meaningful Transparency in Civic AI Systems. *arXiv preprint arXiv:2510.07889* (2025).

[79] Renee Noortman, Britta F Schulte, Paul Marshall, Saskia Bakker, and Anna L Cox. 2019. HawkEye-deploying a design fiction probe. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[80] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. 2021. "Facebook Promotes More Harassment": Social Media Ecosystem, Skill and Marginalized Hijra Identity in Bangladesh. 5, CSCW1, Article 157 (apr 2021), 35 pages. doi:10.1145/3449231

[81] NquiringMinds Ltd. 2025. *Investigation Into X's Use of Personal Data for AI Training Under GDPR.* https://nquiringminds.com/ai-legal-news/investigation-into-xs-use-of-personal-data-for-ai-training-under-gdpr/ Accessed 2025-12-03.

[82] OECD. 2025. *Mapping relevant data collection mechanisms for AI training.* Technical Report No. 48. OECD Publishing. doi:10.1787/3264cd4c-en Accessed 2025-12-03.

[83] Tolulope Oshinowo, Sohyeon Hwang, Amy X Zhang, and Andrés Monroy-Hernández. 2025. Seeing the Politics of Decentralized Social Media Protocols. *arXiv preprint arXiv:2505.22962* (2025).

[84] Christina A Pan, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. 2022. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–31.

[85] Henrik Skaug Sætra and Jo Ese. 2023. Shinigami eyes and social media labeling as a technology for self-care. *Technology and sustainble development: The promise and pittfals of techno-solutionism* (2023), 53–69.

[86] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.

[87] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[88] Giovanni Sartor and Francesca Lagioia. 2020. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.* Technical Report PE 641.530. European Parliamentary Research Service (EPRS), Panel for the Future of Science and Technology (STOA). STOA study.

[89] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-computer Interaction* 2, CSCW (2018), 1–27.

[90] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*. 1–10.

[91] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New media & society* 21, 7 (2019), 1417–1443.

[92] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[93] Carey Shenkman, Dhanaraj Thakur, and Emma Llansó. 2021. Do you see what I see? Capabilities and limits of automated multimedia content analysis. *arXiv preprint arXiv:2201.11105* (2021).

[94] Clare Southerton, Daniel Marshall, Peter Aggleton, Mary Lou Rasmussen, and Rob Cover. 2021. Restricted modes: Social media, content classification and LGBTQ sexual citizenship. *New Media & Society* 23, 5 (2021), 920–938. doi:10.1177/1461444820904362

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

[96] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 20539517221115189.

[97] Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G Deepalakshmi, Jaehyuk Cho, and G Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal* 80 (2023), 110–121.

[98] Melissa Terras, Bettina Anzinger, Paul Gooding, Günter Mühlberger, Michaela Prien, Joe Nockels, C Annemieke Romein, Andy Stauder, and Florian Stauder. 2025. The artificial intelligence cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for automated text recognition. *Open Research Europe* 5 (2025), 16.

[99] Hibby Thach, Samuel Mayworm, Michaelanne Thomas, and Oliver L Haimson. 2024. Trans-centered moderation: Trans technology creators and centering transness in platform and community governance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 326–336.

[100] David Thiel and Renee DiResta. 2023. Addressing Child Exploitation on Federated Social Media. https://cyber.fsi.stanford.edu/io/news/addressing-child-exploitation-federated-social-media

[101] Eddie L Ungless, Nina Markl, and Björn Ross. 2025. Experiences of censorship on TikTok across marginalised identities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 1952–1965.

[102] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. " At the End of the Day Facebook Does What ItWants" How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on human-computer interaction* 4, CSCW2 (2020), 1–22.

[103] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability for content moderation. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–28.

[104] Emily A Vogels. 2021. The state of online harassment. *Pew Research Center* 13 (2021), 625.

[105] Stanislav Vojíř, Zdeněk Smutnỳ, and Jan Kučera. 2020. Social and Technical Aspects of Re-decentralized web. *IDIMT-2020 Digitalized Economy, Society and Information Management* (2020), 107–116.

[106] Ben Wagner, Johanne Kübler, Eliška Pírková, Rita Gsenger, and Carolina Ferro. 2021. REIMAGINING CONTENT MODERATION AND SAFEGUARDING FUNDAMENTAL RIGHTS. (2021).

[107] Owen Xingjian Zhang, Sohyeon Hwang, Yuhan Liu, Manoel Horta Ribeiro, and Andrés Monroy-Hernández. 2025. Understanding Community-Level Blocklists in Decentralized Social Media. *arXiv preprint arXiv:2506.05522* (2025).

[108] Zhilin Zhang, Jun Zhao, Ge Wang, Samantha-Kaye Johnston, George Chalhoub, Tala Ross, Diyi Liu, Claudine Tinsman, Rui Zhao, Max Van Kleek, et al. 2024. Trouble in Paradise? Understanding Mastodon Admin's Motivations, Experiences, and Challenges Running Decentralised Social Media. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–24.

[109] Haris Bin Zia, Aravindh Raman, Ignacio Castro, and Gareth Tyson. 2025. Collaborative Content Moderation in the Fediverse. *arXiv preprint arXiv:2501.05871* (2025).

[110] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 493–502.

[111] Shoshana Zuboff. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019.* Profile books.

[112] Diana Zulli, Miao Liu, and Robert Gehl. 2020. Rethinking the "social" in "social media": Insights into topology, abstraction, and scale on the Mastodon social network. *New Media & Society* 22, 7 (2020), 1188–1205.