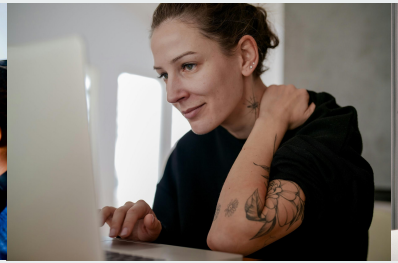
IFTAS

2025 Social Web Trust & Safety Needs Assessment Report





Independent Federated Trust and Safety is a 501c3 organisation, committed to fostering a safe and inclusive open social web

About the Needs Assessment

IFTAS conducts an annual survey of moderators, community managers, and service administrators. For 2025, our participant cohort represents over 7,000,000 hosted accounts on 184 ActivityPub services and communities

As in 2024, we received responses from moderators and community managers on Bluesky and ATProto services, and other Social Web platforms

We also received responses from participants moderating on platforms such as Discord and Reddit. Although we do not include their community membership counts in our totals, their personal experiences provide valuable insight into moderation challenges across a wide range of decentralised and centralised environments

About the Needs Assessment

All answers are optional. When we show percentages, this is the percentage of those who responded. Most questions were answered by 90% or more of all participants

We show individual-level and community-level data. Some responses are from multiple moderators that work on the same service or community as a team

Community-level findings are limited to one response per service or community in this report

Key Findings

Moderator workload is increasing, the mod-to-member ratio has changed from 1 per 1,200 active accounts to 1 per 3,500. This could be due to moderator attrition, growth in accounts, or data from different-sized services

Moderation tools are increasingly designed around the operational realities of large “flagship” communities, creating friction and reduced effectiveness for small communities and single-user instances

Moderator and Admin burnout is a persistent issue - 1 in 5 moderators report experiencing trauma or burnout, underlining the need for wellness and resilience resources

Spam is now the number one issue Admins and Moderators want help dealing with, with CSAM moving to number two

Respondent Profiles

Staff Role

Multiple choice, respondent can select one or more

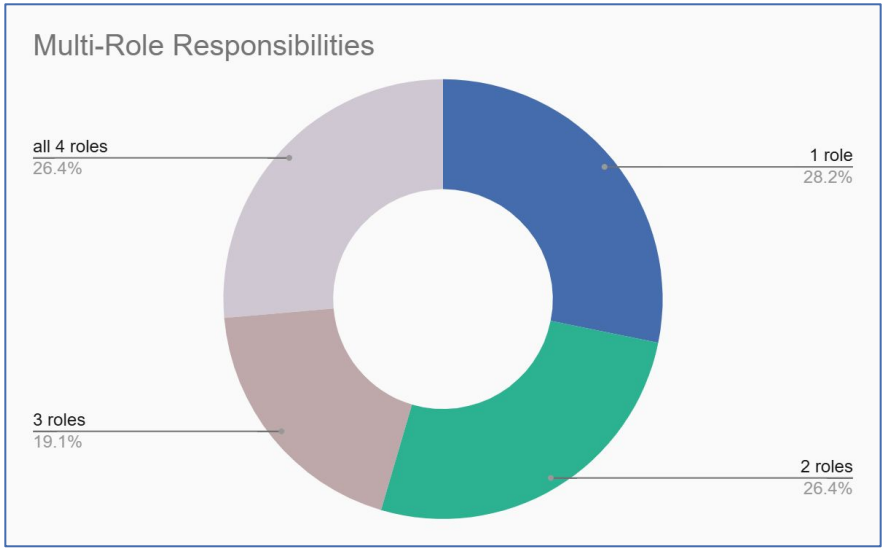
Many selected all four roles, suggesting a high workload

Role overlap increases pressure on individuals. Support systems should be flexible and broadly applicable

Administrator, Moderator, Community Manager, Legal	29
Administrator, Moderator, Community Manager	19
Administrator, Moderator	18
Administrator	17
Moderator	12
Moderator, Community Manager	6
Community Manager	2
Administrator, Community Manager	2
Administrator, Moderator, Legal	2
Administrator, Legal	2
Moderator, Legal	1

Staff Role


What is your role?
100
75
50
25
0
Community Manager
Content Moderator
Legal / Compliance
Service Administrator

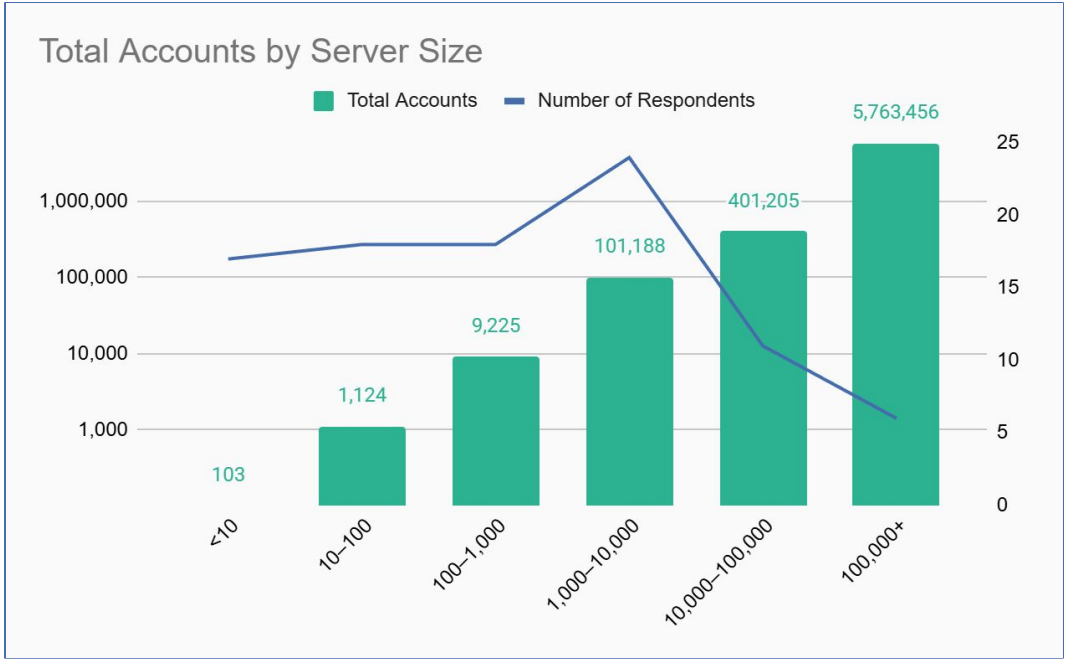

Multi-Role Responsibilities
all 4 roles
26.4%
1 role
28.2%
3 roles
19.1%
2 roles
26.4%

Service Provision

Respondents host a collective 7,043,703 accounts - roughly 42% of all known ActivityPub accounts

76% of respondents host 10,000 accounts or fewer; 47% under 100

Tooling and resources designed for larger services may be misaligned with the reality of how most respondents operate

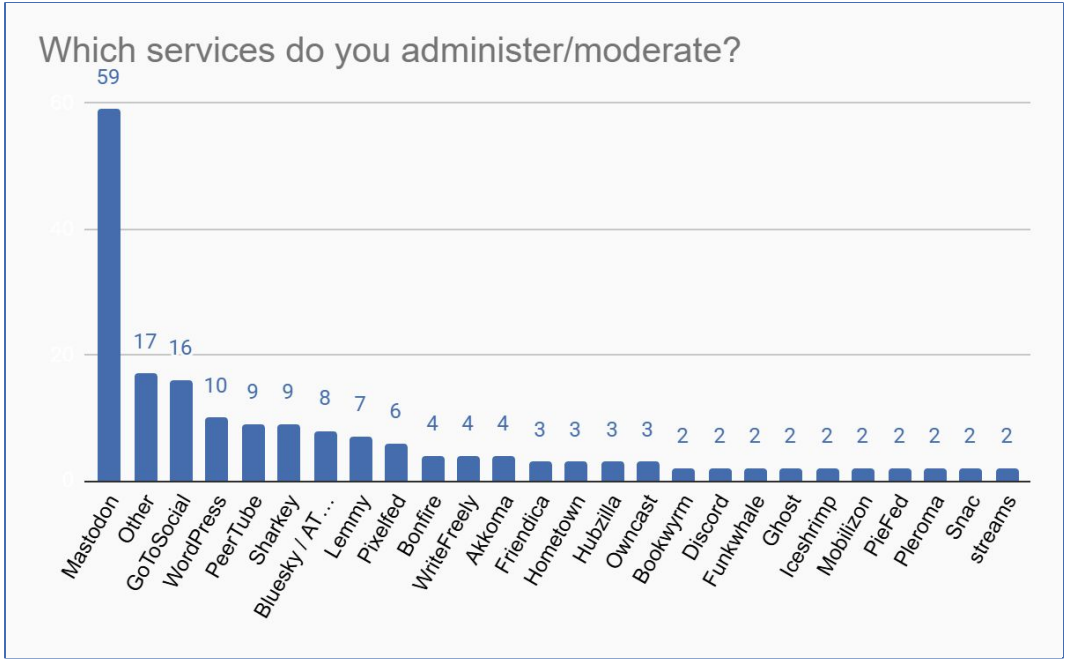


Number of Services

184 individual services reported

32% are Mastodon servers

Over 40 distinct platforms reported, with many respondents active across multiple platforms

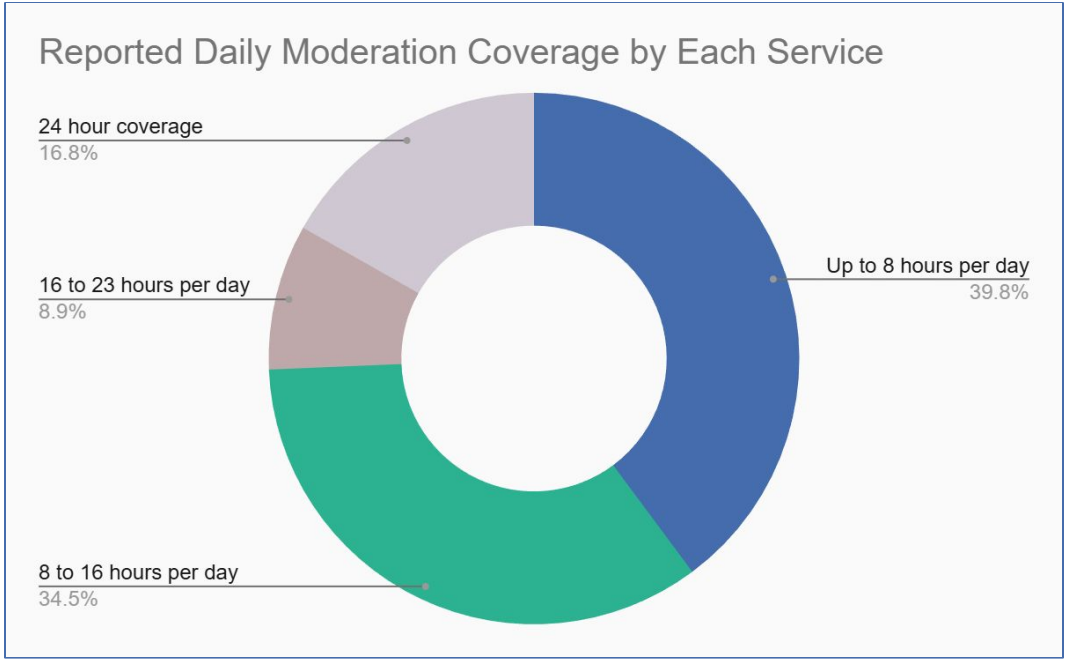


Staff Coverage

Respondents reported a total of 290 moderators across these 184 communities

16.8% of services provide moderation 24 hours a day

Roughly 1 Moderator for every 3,500 active accounts (down from 1:1,200 last year)

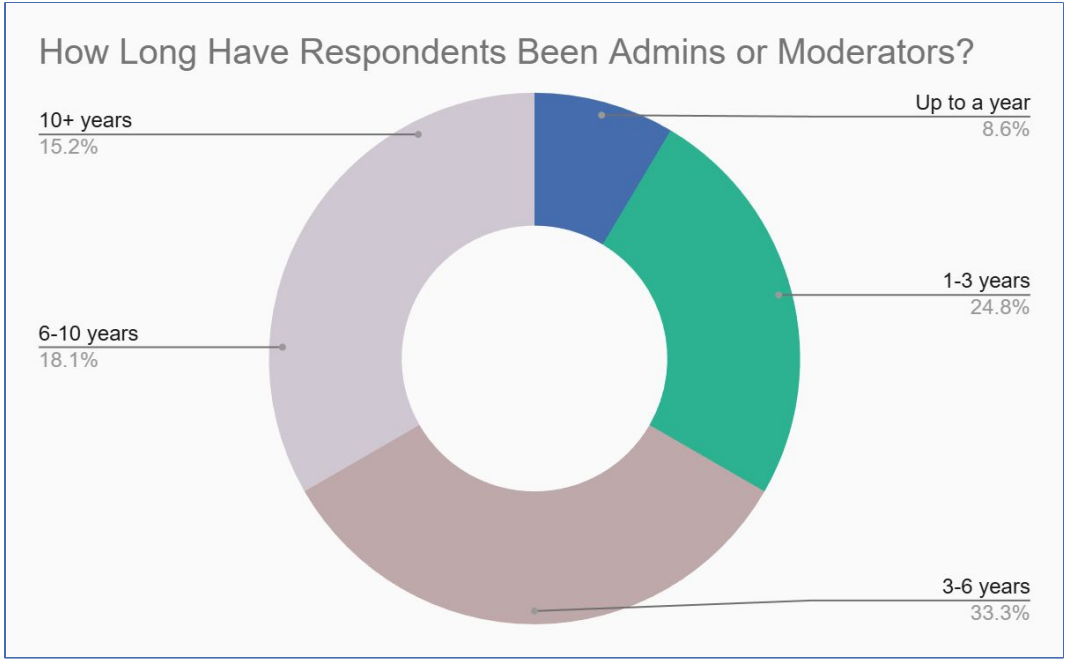


Experience

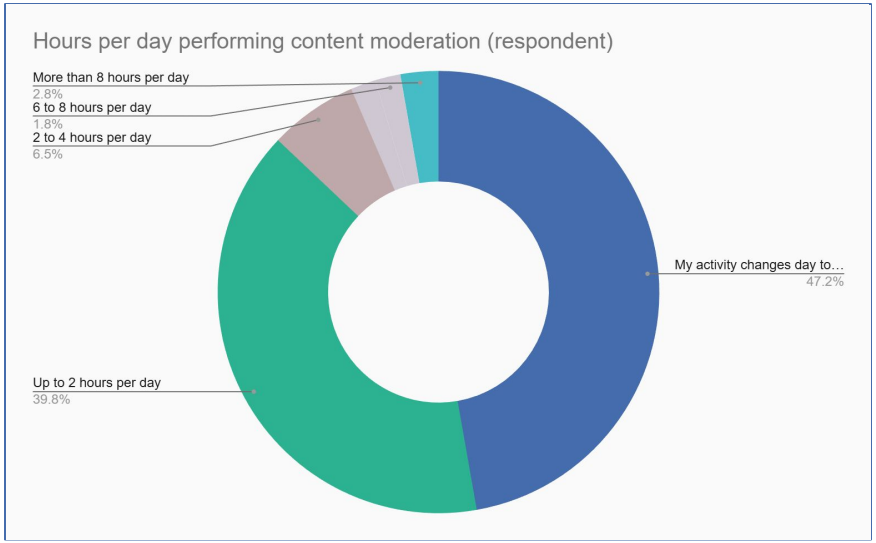
Nearly 60% of moderators have between 1 and 6 years of experience

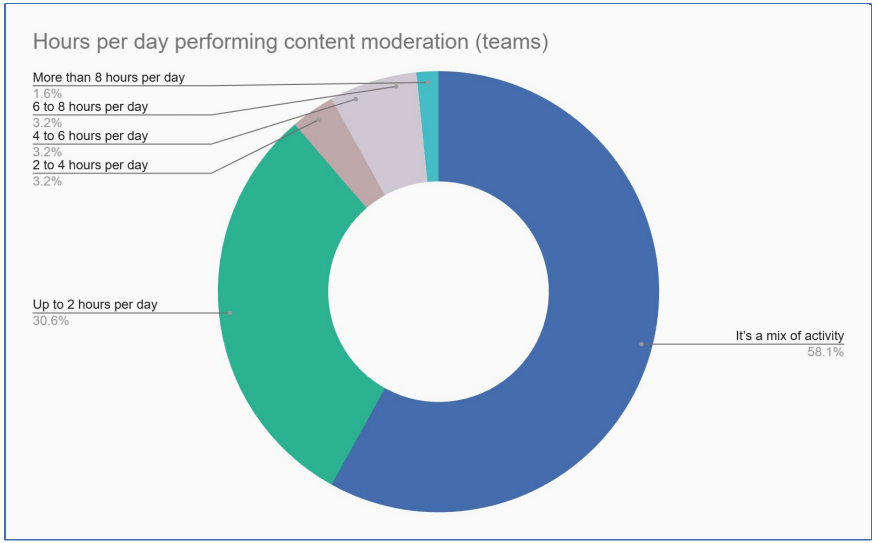
8.6% are new to moderation this year - down from 18.4% in 2024

Overall, experience is rising, but reduced onboarding may pose sustainability risks

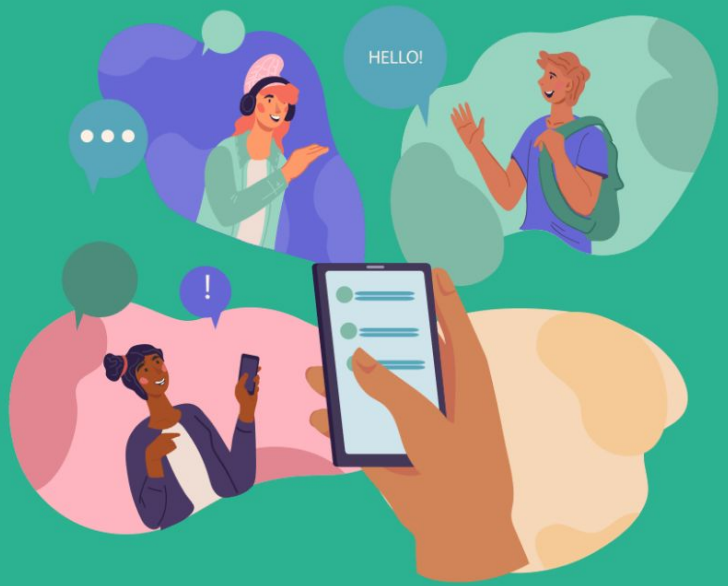


Workload


Hours per day performing content moderation (respondent)
More than 8 hours per day
2.8%
6 to 8 hours per day
1.8%
2 to 4 hours per day
6.5%
My activity changes day to…
47.2%
Up to 2 hours per day
39.8%


Hours per day performing content moderation (teams)
More than 8 hours per day
1.6%
6 to 8 hours per day
3.2%
4 to 6 hours per day
3.2%
2 to 4 hours per day
3.2%
Up to 2 hours per day
30.6%
It's a mix of activity
58.1%

Moderator Support



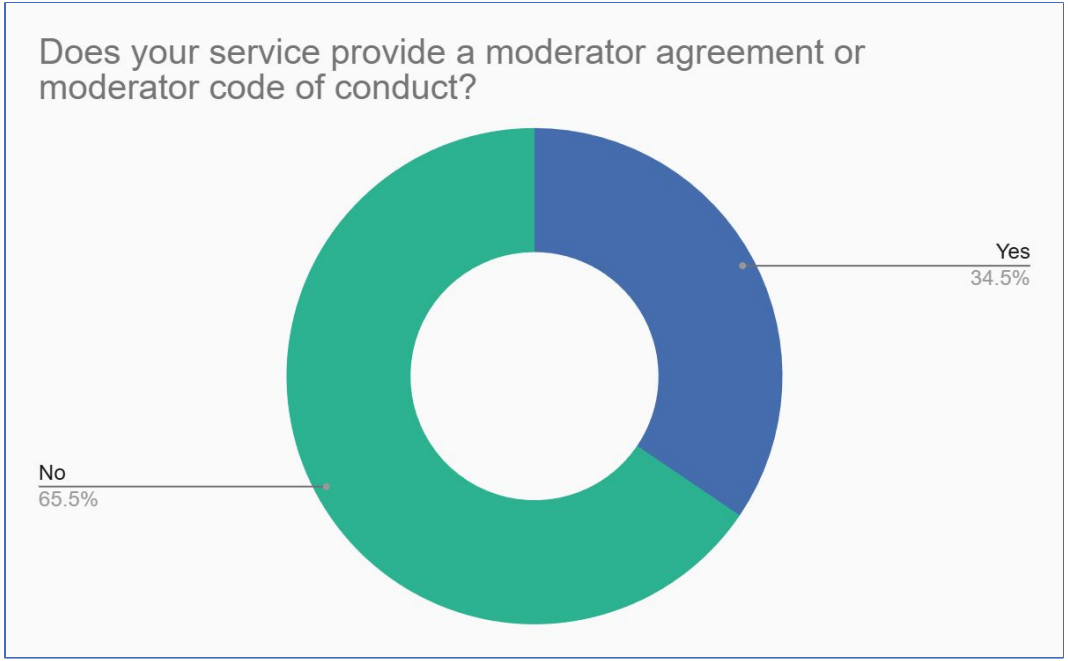
Moderator Agreement

Solo moderator services excluded

50% of respondents reported a team of two or more moderators

Two thirds of these teams operate without a formal moderator agreement defining roles and responsibilities

✅ *Sample moderator agreements are available at https://about.iftas.org/library/sample-moderator-agreements/*



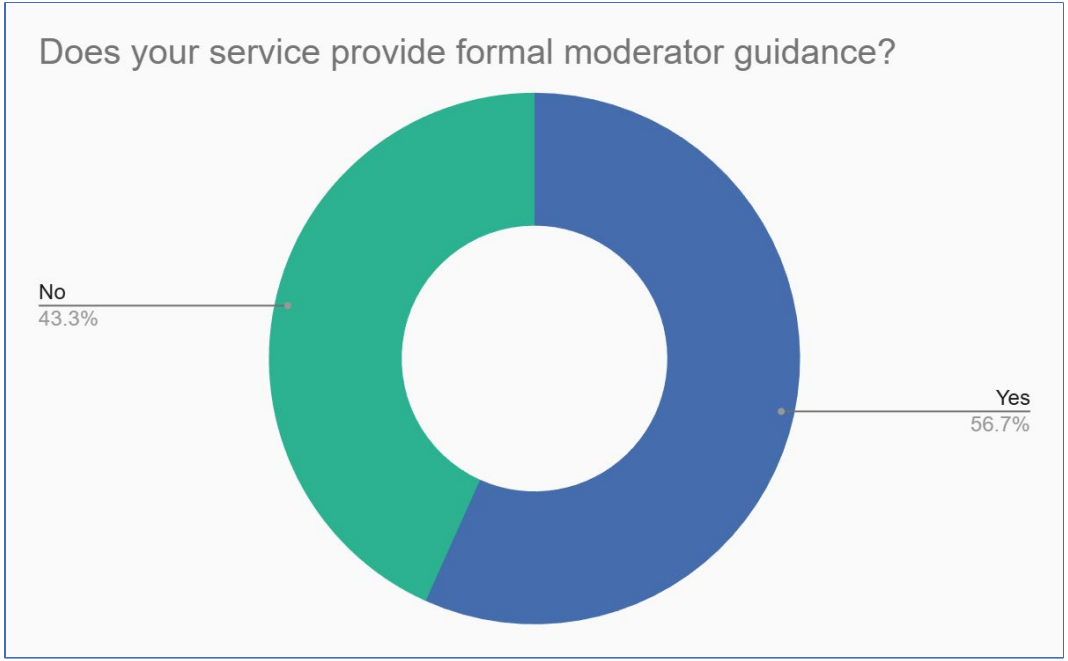
Moderator Guidance

Solo moderator services excluded

Over 40% of respondents with two or more moderators do not provide formal moderator guidance

Moderation decisions may rely on informal norms rather than shared standards

✅ *Moderator guidance can be found at https://about.iftas.org/library/content-moderation-educational-resources/*

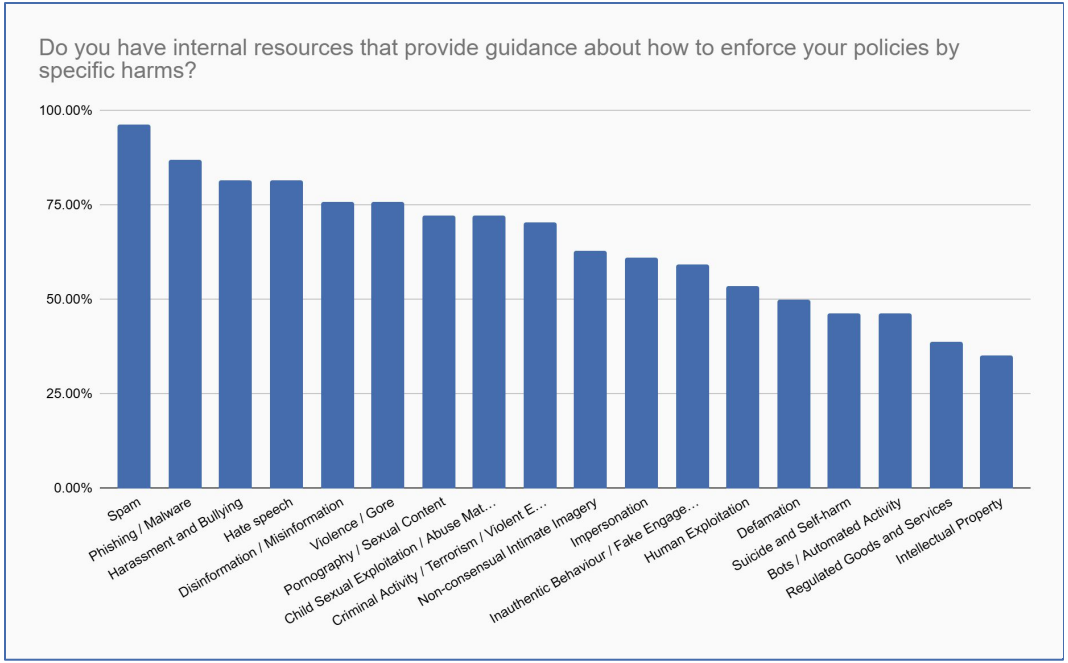


Moderator Guidance

Communities that do have policy enforcement guidance reported if that guidance covers specific categories of harm

Policies often cover select harm categories rather than the full spectrum

✅ *Definitions of harms and associated guidance is available at https://about.iftas.org/wiki_cats/content/*



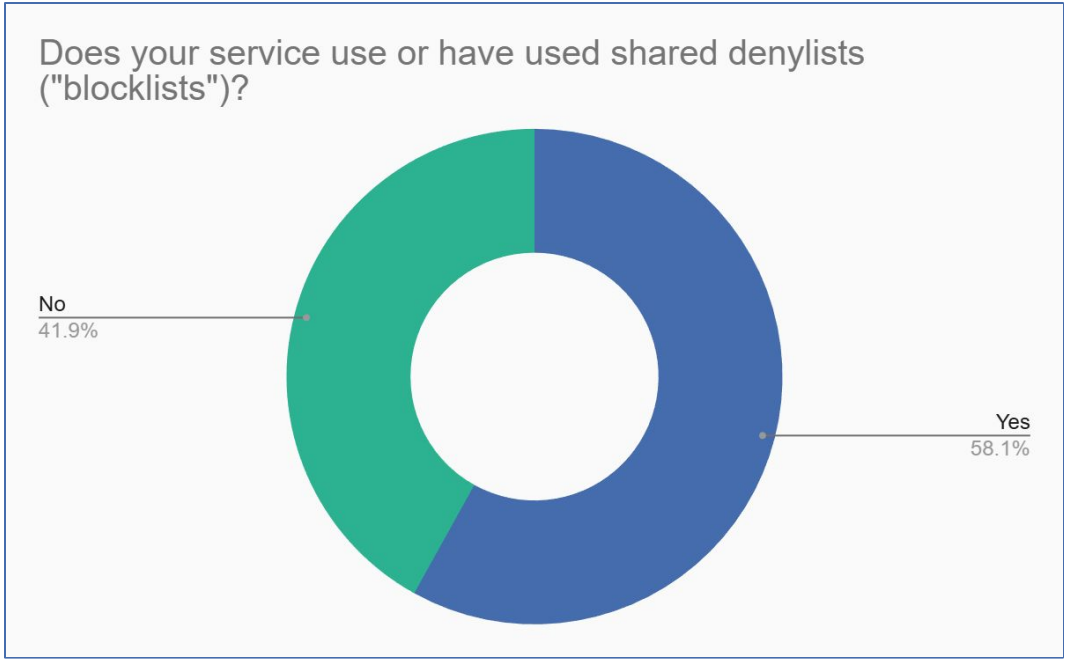
Domain Denylisting

Increase from 50% in 2024 to 58%

Commonly used sources include:

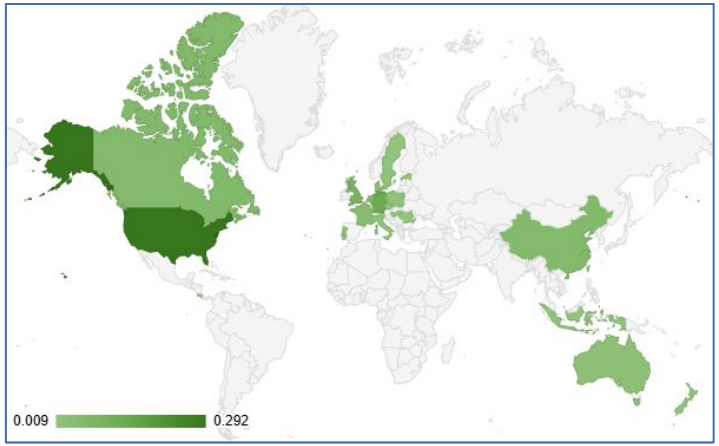
- IFTAS
- Oliphant Lists
- Seirdy Lists
- #Fediblock (hashtag)
- Fediseer
- Gardenfence

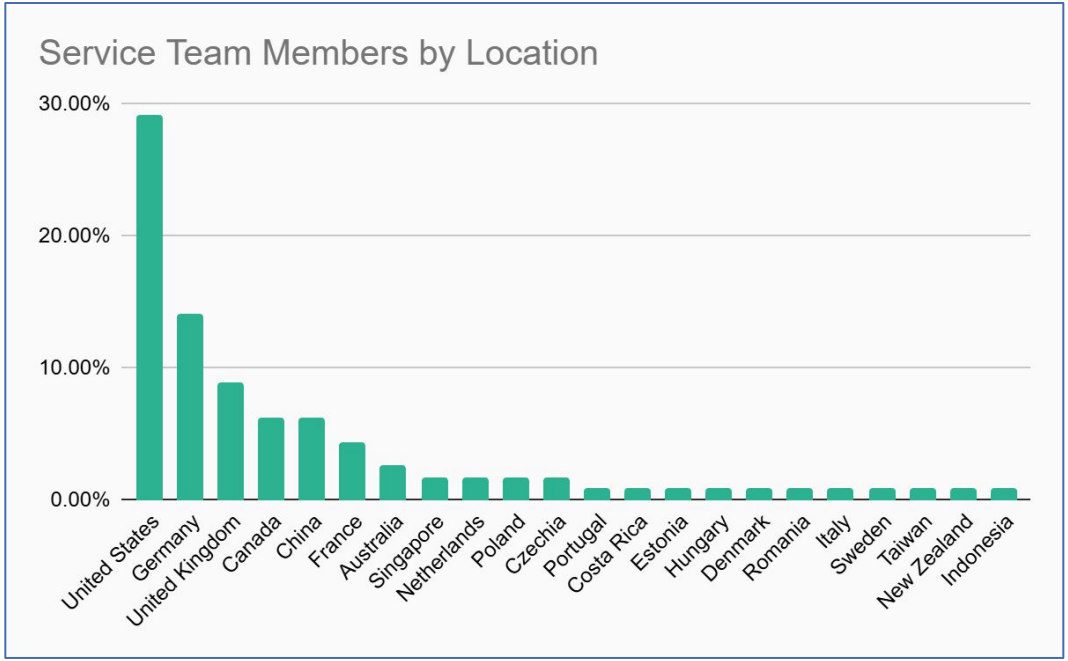
✅ *Denylist resources are available at https://about.iftas.org/library/denylist-resources/*



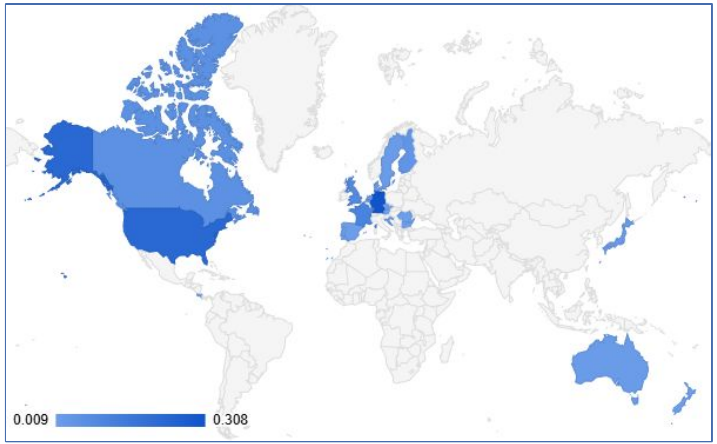
Organisational Status

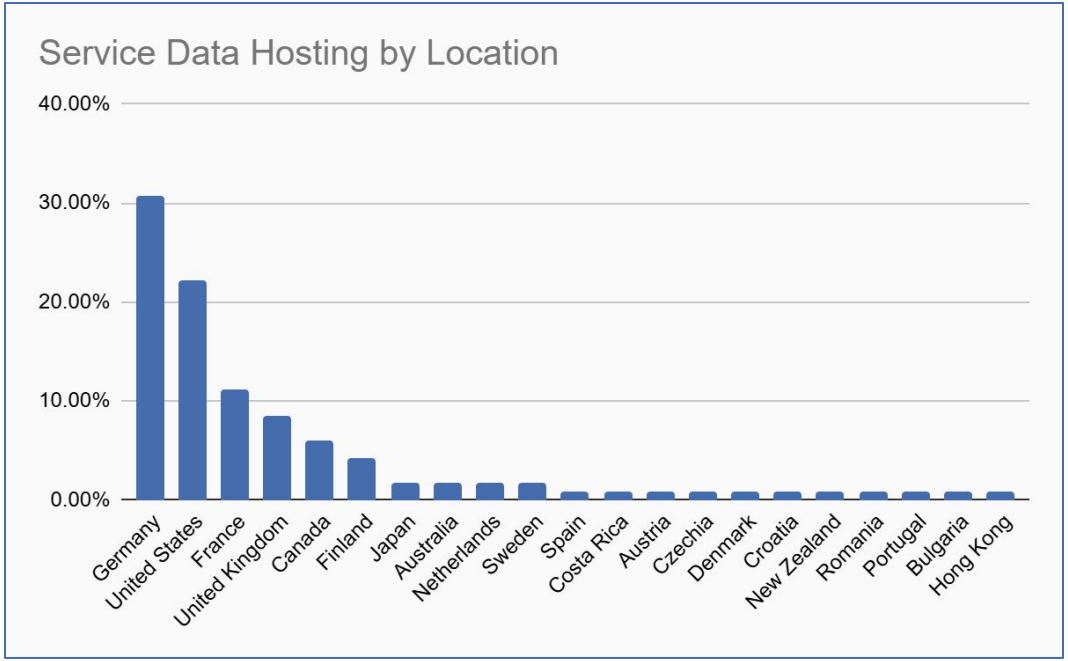
Staff by Location





Data Storage by Location





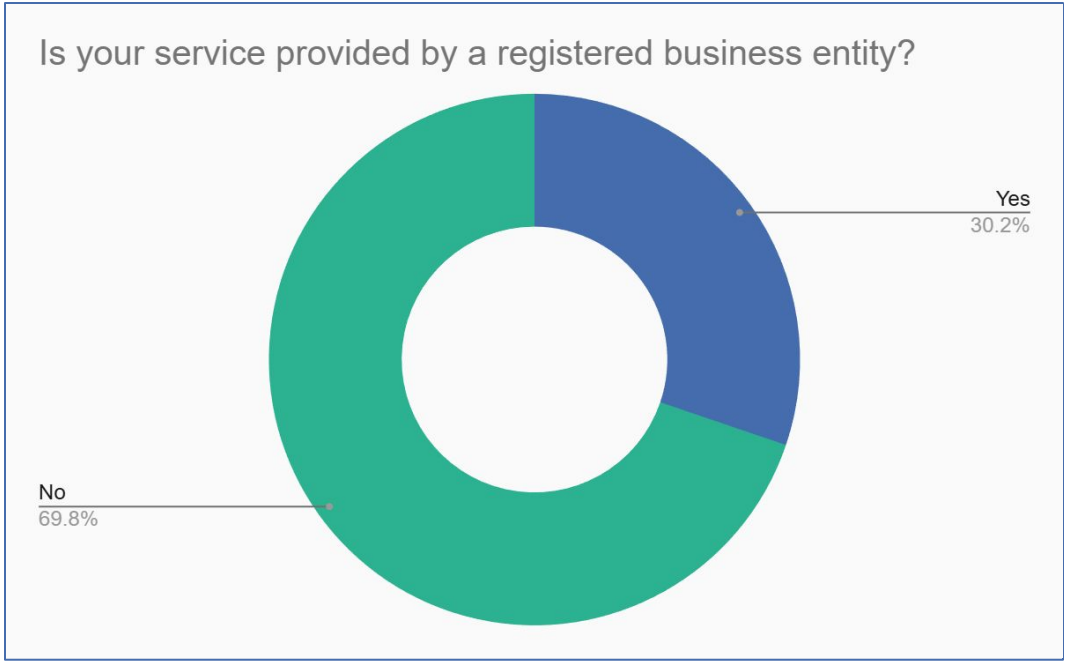
Business Operations

30% of respondents are moderating for a registered business (up from 22%)

Only 5% of these provide business or liability insurance coverage

Majority of registered businesses are in the United States

Non-profits and small companies dominate

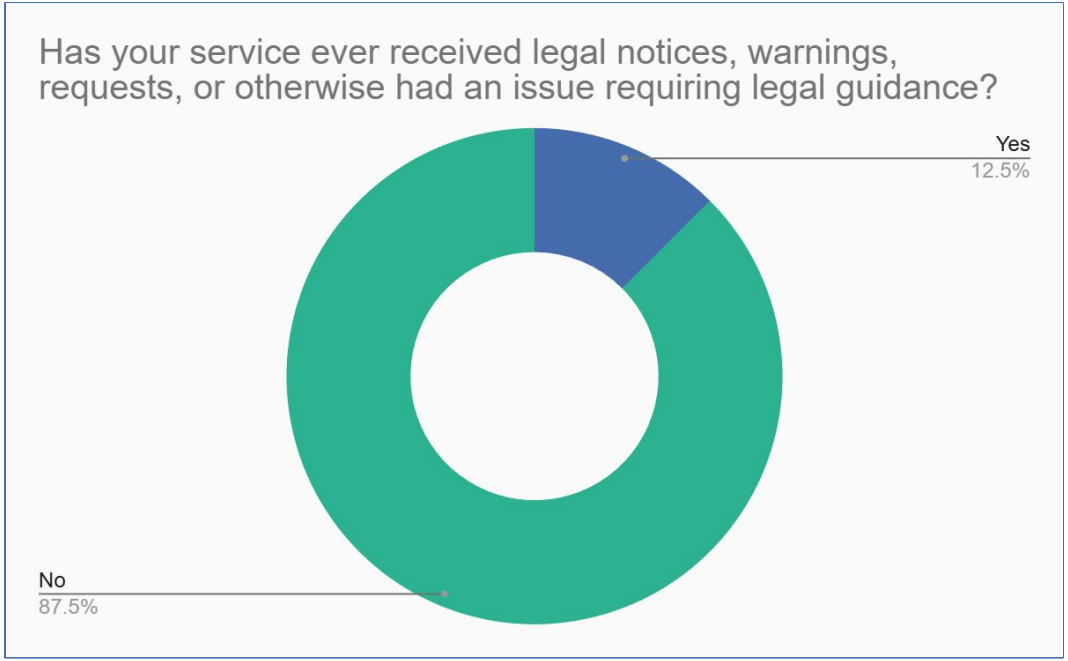


Business Operations

Most common issues include copyright takedowns, CSAM-related incidents, and GDPR-related threats

A few services faced serious legal interactions, including FBI requests and libel complaints, highlighting the need for access to legal guidance

✅ *Introductory legal guidance available at https://about.iftas.org/library/legal-background-reading/*

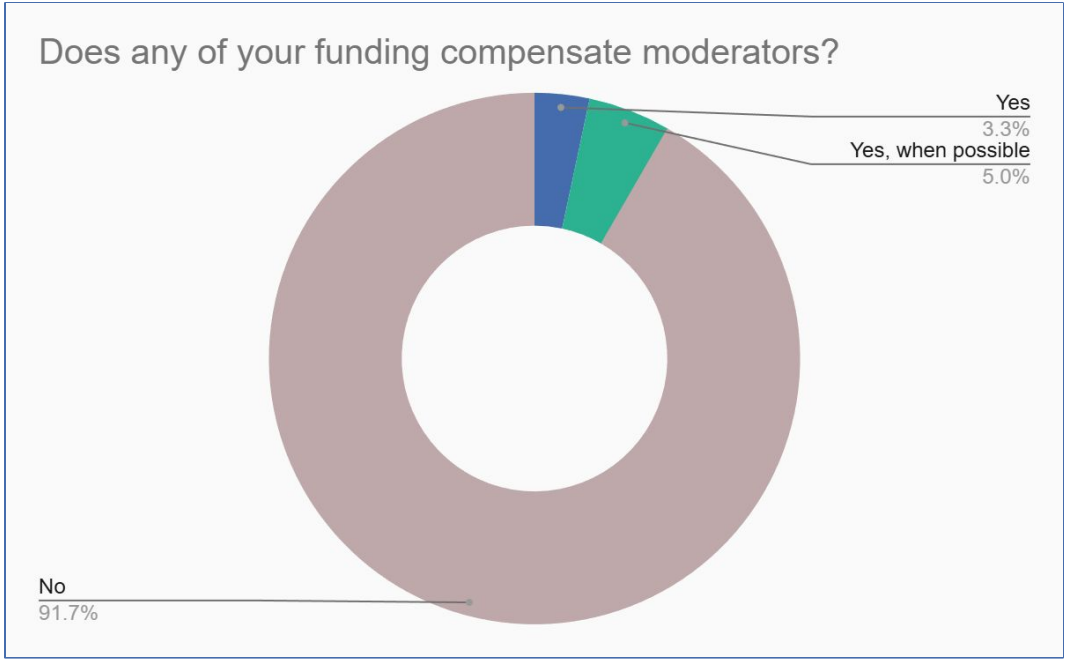


Business Operations

60% of services solicit donations or monetary support of some kind

8% compensate moderators for their labour (if and when possible, usually from excess donations)

Of the respondents that chose to share revenue and expense data, these communities collectively raise $13,933 each month, failing to cover fixed costs of $21,402

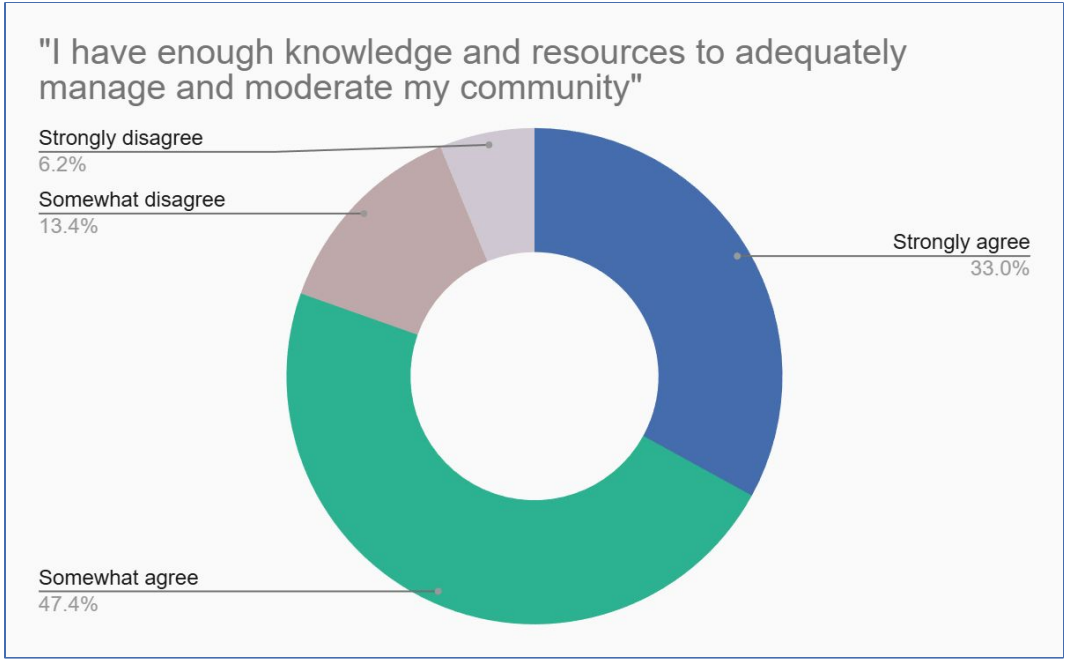


Needs Assessment

Self-assessment

Strongly Disagree up from 2% to 6%

Over one in five expressed some level of concern, highlighting gaps in resources or knowledge



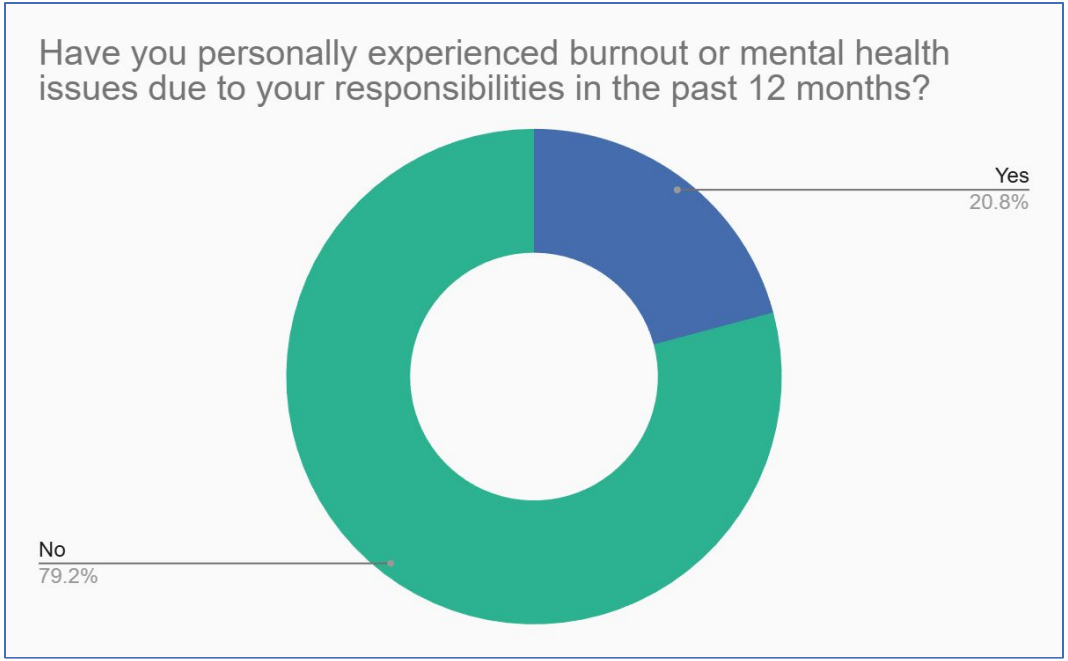
✅ The IFTAS Community Library *is available at* *https://about.iftas.org/trust-safety-services/iftas-community-library/*

Wellness

One in five experiencing trauma or burnout, unchanged since 2023

The persistence of this rate highlights ongoing structural issues: operating without formal training or support, moderating in isolation, exposure to high volumes of distressing content

✅ *Wellness and resilience resources are available at https://about.iftas.org/wiki_cats/wellness/*



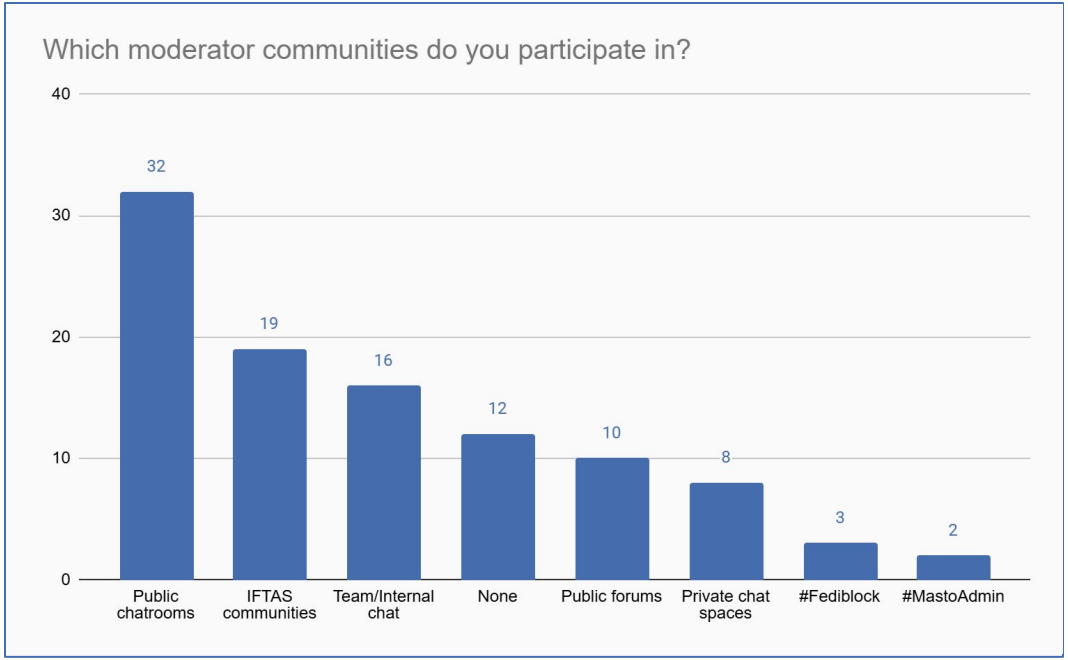
Community Support

Collaboration is informal and real-time, favouring chat over structured or archival spaces

Shared spaces matter, with IFTAS communities functioning as a key connective layer

Many moderators remain isolated, either by choice or lack of access

✅ *A list of communities can be found at https://about.iftas.org/library/technical-support-communities/*



Resource Needs

Participants were asked to stack rank **resource** needs

Regulatory guidance climbing, likely due to increased compliance requirements from UK, Australia, US

Shared denylists added to this section, entered the list at number three

Consent-based federation tooling (e.g. allowlisting, trust indicators) emerged strongly in qualitative responses but was not included in the original survey ranking options

✅ The IFTAS Community Library *is available at* https://about.iftas.org/trust-safety-services/iftas-community-library/

1 ↑ (1)	Regulatory guidance (GDPR, OSA etc)
2 ↓ (1)	Moderation guidance / best practices
3 (new)	Shared denylists / allowlists / blocklists
4 ↓ (1)	Template community guidance / code of conduct
5 →	Help with responding to content takedowns, law enforcement, legal notices
6 ↓ (2)	Template legal documents
7 ↓ (1)	Moderator training courses
8 →	Moderator wellbeing resources
9 →	Resources in languages other than English
10 →	Help with business formation

Tools and Services Needs

Participants were asked to stack rank **tooling or service** needs

Spam detection, and spam IP/email domains moving up

Disinformation climbs three places this year, hate speech up two

✅ The IFTAS Community Library *is available at* https://about.iftas.org/trust-safety-services/iftas-community-library/

1 ↑ (1)	Spam detection
2 ↓ (1)	CSAM detection
3 ↑ (2)	Toxic / spam IP address and email domain data
4 ↑ (2)	Hate speech detection
5 ↓ (1)	Phishing and malware detection
6 ↓ (3)	Non-consensual intimate image detection
7 ↑ (3)	Disinformation detection
8 ↓ (1)	Personal digital safety tools
9 ↓ (1)	Terroristic and violent extremism detection
10 (new)	Registered agent / point of contact service

Impact Assessment

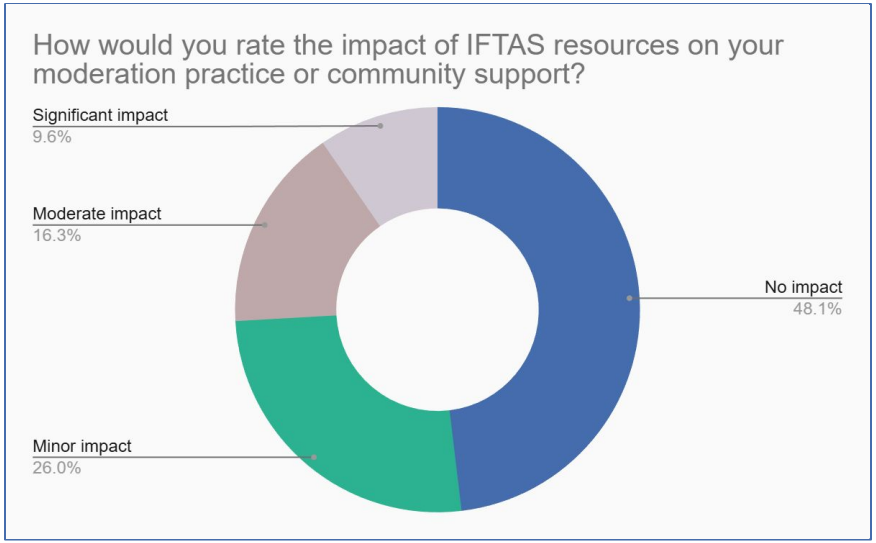
Assessing IFTAS Support

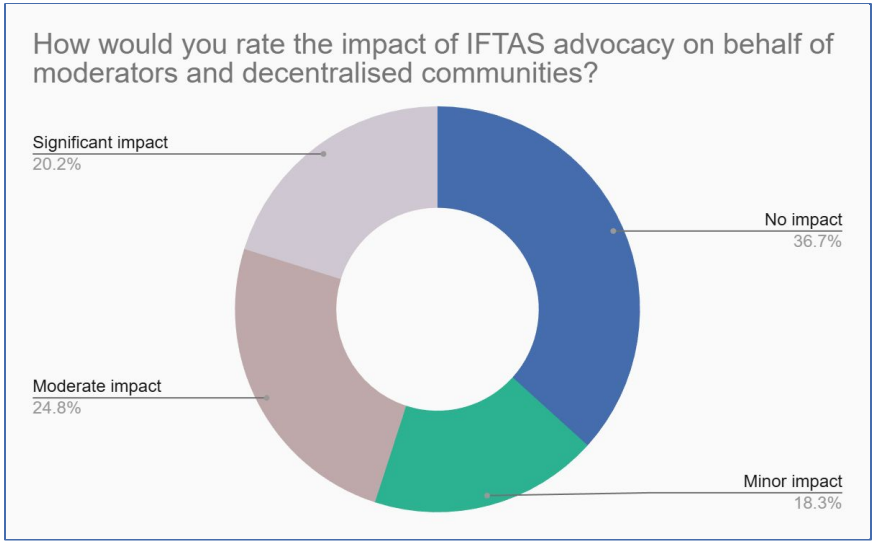
For the first time, this year's survey asked moderators and administrators to assess the usefulness of IFTAS resources and support

Respondents rated written materials, tools, advocacy efforts, and connection to peers

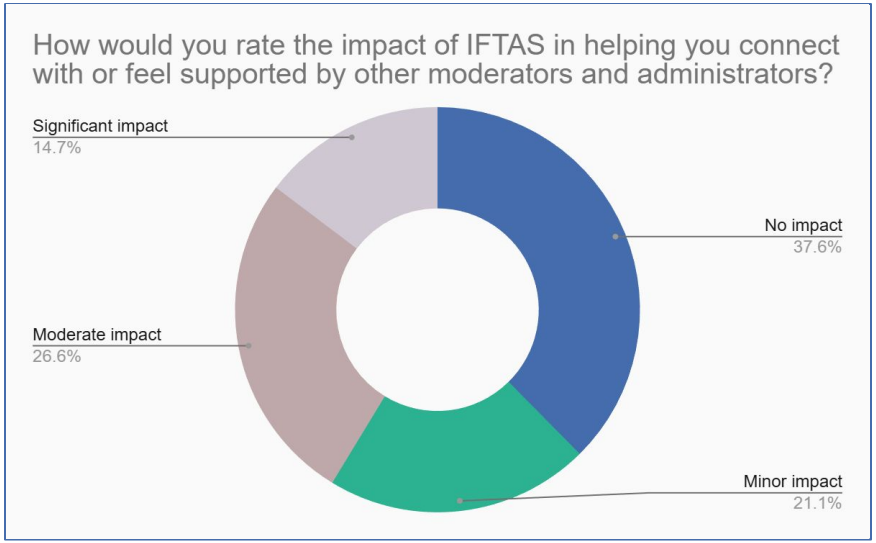
This feedback gives us a baseline to understand where IFTAS is delivering value, and where gaps remain

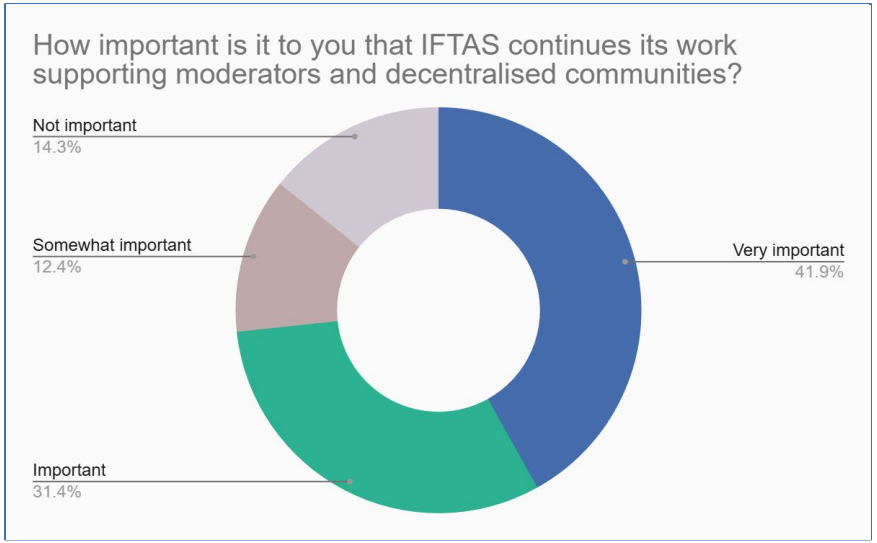
IFTAS Impact





IFTAS Impact





Assessing IFTAS Support

Mixed impact overall, most respondents saw at least minor benefit, but few rated resources as highly impactful. Advocacy and community support rated more positively than static resources

Connection-building efforts are valued but not reaching everyone, some respondents still feel isolated or unsupported

The data reflects a clear appetite for continued and improved support, stronger tooling, and better integration with the lived experience of moderators across platforms

There is a strong foundation of goodwill and perceived potential, IFTAS is seen as an important actor in this space, even by those who haven't yet directly benefited from its work

Key Needs

Key Needs and Issues

- Spam, CSAM, hate speech, and disinformation detection tools are top priorities, but access to high-quality, affordable, and accurate tools remains inconsistent
- Services regularly face copyright takedowns, GDPR challenges, and CSAM-related incidents. A few even report FBI requests and libel threats, but most lack access to legal guidance or response frameworks
- Many moderators rely on informal group chats or manually maintained denylists. Automation, interoperability between tools, and transparency remain critical gaps

Key Needs and Issues

- Burnout is persistent. Nearly 1 in 5 moderators continues to experience burnout, trauma, or overwhelm, with no significant improvement since 2023, indicating structural issues around workload, isolation, and lack of support
- 47% of participants manage communities of fewer than 100 accounts. Tooling and support structures still often prioritise large platforms, leaving smaller communities underserved
- Increasing concerns over growing numbers of Internet safety regulations that impact day-to-day operations, including international/extraterritorial compliance issues

Key Needs and Issues

Note: After the survey closed in September 2025, a coordinated pro-Russian propaganda campaign targeting ActivityPub and ATProto services emerged in October, creating hundreds of accounts connected to automated disinformation feeds.

While not reflected in survey responses, this development underscores the evolving threat landscape and highlights the continued importance of cross-network moderation strategies.

(Learn more)

Looking Ahead

2026 Forecast: Collaborative Moderation & Knowledge Sharing

Current moderation is reactive and isolated, with most teams relying on manually maintained denylists rather than shared behavioural signals

Account holders face a fragmented reality where the same content is permitted on one service but leads to defederation on another, creating friction for the open social web

In 2026, the need will shift from simple denylists to shared moderation logic. Communities will seek ways to exchange trust signals and automated classifier insights without compromising local autonomy

2026 Forecast: Evolving Threat Vectors

Spam has moved to the number one issue, but it is rapidly evolving into automated disinformation feeds and deepfake-based impersonation

As seen with the "Portal Kombat" network in late 2025, pro-Russian propaganda and other state-actor campaigns are now targeting ActivityPub and ATProto simultaneously

Moderators will require detection tools for video and audio (eg Peertube/Owncast/Loops) as automated harms move away from easily filtered text into more complex, "human-mimicking" media

2026 Forecast: Discerning Authenticity in the AI Flood

In 2026, the gap between human and synthetic content is projected to become nearly invisible as AI creators become emotionally convincing and culturally aware

Human moderators work within real-world constraints, while AI systems can produce high-volume, optimised content with zero fatigue, leading to a content flood that threatens to overwhelm volunteer moderation efforts

Audiences are becoming increasingly skeptical of *all* digital content; without clear signals of authenticity, engagement will weaken as participants become too cautious to connect with what they see

Moderators will require content provenance frameworks (like C2PA) to separate human content from machine-made

2026 Forecast: Infrastructure Pressure & Commercial Capture

The presence of multi-million account services shifts the mod-to-member ratio even further, currently at 1 per 3,500 active accounts

While 47% of participants manage communities of fewer than 100 accounts, they must now defend against automated threats designed to target large platforms

Moderation teams will face commercial capture risks, where standardising on a few corporate-backed moderation APIs and data services becomes the only viable way to handle high-volume traffic, potentially undermining decentralisation

2026 Forecast: Regulatory Pressures & Enforcement Mandates

2026 is expected to be a consequential year for enforcement of global digital governance laws, requiring platforms to have more robust policies and audit trails

Small community administrators increasingly face legal notice risks from jurisdictions outside their home country

A critical need for Registered Agent services and template legal frameworks will emerge as a baseline for community survival, as regulators move away from voluntary safety codes toward mandatory, enforceable standards

2026 Forecast: The Rise of Consent-Based Federation

The Problem

Today's default model assumes every service connects to every other by default - communities must actively block harmful or hostile instances after-the-fact

Denylisting is reactive, labour-intensive, and unevenly maintained

Small communities are overwhelmed, forced to adopt blocklists they can't vet

New threats like disinformation networks and AI spam scale faster than human moderation can respond

The Path Forward

Greylisting or Allowlisting by default: New servers do not auto-federate - they're discoverable, but not connected until explicitly approved

Communities choose when and how to federate based on their values, or the values of other communities they trust

Moderators will access trust signals, community health indicators, and social contracts

Federation becomes intentional, not accidental - enabling safer growth, stronger norms, and resilience

Future Outlook: IFTAS in 2026

- IFTAS is critically underfunded, currently operating with one full-time unpaid staff member
- Shifting focus to advocacy, coordination - no longer able to operate some services

What we're still doing:

- SW-ISAC continues to publish alerts and share threat intelligence
- DNI and AUD denylists are still maintained
- Community Library, Domain Observatory, and Signal channels remain online and active

All our available resources can be found at https://about.iftas.org

The Social Web at a Crossroads

2025: Where We Stand

- Volunteer moderation is strained: 1 moderator per 3,500 accounts
- Spam now the top threat, displacing CSAM
- Burnout is chronic: 1 in 5 moderators report trauma or exhaustion
- Small communities dominate: 76% host under 10,000 accounts; 47% under 100
- Most teams lack formal policies, legal help, or training, with legal and safety pressures rising, including extraterritorial risks

2026: What Lies Ahead

- Moderation must move beyond manual, reactive denylisting
- Deepfakes and synthetic content will blur authenticity
- Few tools, controlled by few actors, threaten decentralisation
- Small teams must counter threats built for large platforms
- Global regulation will demand legal readiness and audit trails
- Consent-based federation will become essential

Thank you for participating!

Participant Comments

"A universal cross-platform spam detection tool... An instance administrator can write rules that rapidly deploy across ten thousand instances"

"CSAM classifier models are a need"

"We need collaboration tools between instances. A way to share and receive moderation decisions between instances we trust"

"It would be also nice to see any resources specifically for single-tenant or low-tenant instances"

"Shared denylists, external moderation support, or shared resource pools such as CDNs"

Concerns Raised

“I think the [CSAM scanning] services are a data grab”

“I think you exist to marginalize neuroatypicals”

“you are building mechanisms for government/corporate control”

"I do not find your services useful”

Our Commitments

We received a small number of critical comments this year, including concerns that content classification tools may compromise privacy, or that our work risks enabling centralised or corporate control of the Social Web

We do not and will not support tools that enforce conformity or "normalcy" as a condition for participation

Everything we provide or share is opt-in, transparent, and designed with decentralisation in mind. We recognise not every instance or admin will need the same level of support. Our goal is to be there if and when support is wanted

Messages of Support

“Thanks for your effort and supporting the fediverse!”

“Really appreciate the work you all do, thanks!”

"You do great work and I hope you all are able to get funding as the work you do is critical to the functioning of the Fediverse”

“The templates have been very useful”

“Keep going! For a better, free and decentralized Internet!”

Help Sustain IFTAS

IFTAS fills critical gaps in trust & safety infrastructure for the decentralised social web. Without support, this work is at risk of disappearing just as adoption and threats are growing

How you can help:

- Fund or sponsor IFTAS' continued work: https://about.iftas.org/donate/
- Join with or advocate for organisations that support decentralised social networks
- Volunteer your expertise
- Help us tell the story - share this report!

Thank You to our Admin & Moderator Community!

Thank you for your time, energy, participation, and for everything you do to keep the Fediverse safe! 💙

To Learn More

- about.iftas.org
- mastodon.iftas.org/@iftas