ANNEX 5

SCALING TRUST ON THE WEB

COLLECTIVE SECURITY IN A FEDERATED WORLD

COMPREHENSIVE REPORT OF THE TASK FORCE FOR A TRUSTWORTHY FUTURE WEB

The mission of the Digital Forensic Research Lab (DFRLab) is to identify, expose, and explain disinformation where and when it occurs using open-source research; to promote objective truth as a foundation of government for and by people; to protect democratic institutions and norms from those who would seek to undermine them in the digital engagement space; to create a new model of expertise adapted for impact and real-world results; and to forge digital resilience at a time when humans are more interconnected than at any point in history, by building the world's leading hub of digital forensic analysts tracking events in governance, technology, and security.

ISBN: 978-1-61977-279-3

This report is written and published in accordance with the Atlantic Council Policy on Intellectual Independence. The authors are solely responsible for its analysis and recommendations. The Atlantic Council and its donors do not determine, nor do they necessarily endorse or advocate for, any of this report's conclusions.

© **2023 The Atlantic Council of the United States.** All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means without permission in writing from the Atlantic Council, except in the case of brief quotations in news articles, critical articles, or reviews.

Please direct inquiries to:
Atlantic Council
1030 15th Street, NW, 12th Floor
Washington, DC 20005

For more information, please visit www.AtlanticCouncil.org

June 2023

William + Flora Hewlett Foundation SCHMIDT FUTURES dūcō

ANNEX 5

COLLECTIVE SECURITY IN A FEDERATED WORLD

TABLE OF CONTENTS

INTRODUCTION

Many discussions about social media governance and trust and safety—among regulators, developers, researchers, and users alike—are focused on a small number of centralized, corporate-owned platforms that currently dominate the social media landscape: Meta's Facebook and Instagram, YouTube, Twitter, Reddit, and a handful of others. The emergence and growth in popularity of federated social media services, like Mastodon and Bluesky, introduces new opportunities, but also significant new risks and complications. While federated services continue to be dwarfed in size in comparison to platforms like Facebook and Twitter, the steady rise in their adoption warrants further attention and study. In the case of Mastodon, for example, changes in ownership and governance at Twitter appear to have significantly accelerated the platform's adoption, with some estimates showing more than ten million currently active users. For all the optimistic rhetoric that Mastodon is "like Twitter, but without the bad parts," we should assume that centralized and decentralized platforms share a common set of threats from motivated malicious users—and require a common set of investments to ensure trustworthy, user-focused outcomes.

Broadly speaking, the "fediverse" is a catch-all term for a wide array of distinct products, services, and platforms that interconnect using a set of shared communication protocols such as the W3C standard ActivityPub or the under-development Bluesky AT Protocol. In place of a centralized social media platform like Twitter, a federated alternative might involve dozens, hundreds, or even thousands of individual servers running instances of an open-source product. Despite being maintained by separate people or groups, servers using the same underlying protocol are interoperable, communicating with each other (and, in turn, allowing their users to access one another's' content). A number of distinct products have been built atop these decentralized standards, including Mastodon (a Twitter-like social media platform) and Pixelfed (an Instagram-like platform focused on media sharing).

These emergent distributed and federated social media platforms offer the promise of alternative governance structures that empower consumers and can help rebuild social media on a foundation of trust. Their decentralized nature enables users to act as hosts or moderators of their own instances, increasing user agency and ownership, and platform interoperability ensures users can engage freely with a wide array of product alternatives without having to sacrifice their content or networks. Unfortunately, they also have many of the same propensities for harmful misuse by malign actors as mainstream platforms like Facebook and Twitter, while possessing few, if any, of the hard-won detection and moderation capabilities necessary to stop them. More troublingly, substantial technological, governance, and financial obstacles hinder efforts to develop these necessary functions.

This paper offers an assessment of the trust and safety (T&S) capabilities of federated platforms—with a particular focus on their ability to address collective security risks like coordinated manipulation and disinformation.[1] We focus on disinformation risks for two reasons. First, they have significant societal impact. Second, disinformation threats primarily are detectable and mitigable as actor- and behavior-level phenomena, rather than the content-level moderation approaches discussed in most research about trust and safety.

Beginning with a broad review of the current structures and practices of moderation on federated services, we examine the particular issues created by persistent, adversarial campaigns. We identify several significant structural impediments to robust mitigation of disinformation threats, given current technical and labor models of moderation: namely, the shortcomings of content-driven approaches to moderation in counteracting these campaigns, and the obstacles to implementing behavioral defenses.

MODERATING THE FEDIVERSE

Most discussions of fediverse moderation have, reasonably, focused on the essential contrast between centralized, corporate approaches to content governance (like those employed by Meta, Google, and Twitter), and a distributed, community-driven approach native to federated services like Mastodon. The essential feature of federated systems, and of the protocols like ActivityPub underlying them, is decentralization. Each instance of a federated service can choose for itself what its governance approach will be; in turn, its governance decisions extend only so far as the (virtual) boundaries of that particular server. As Alan Rozenshtein summarizes, "No instance can control the behavior of any other instance, and there is no central authority that can decide which instances are valid or that can ban a user or a piece of content from the ActivityPub network entirely. As long as someone is willing to host an instance and allow certain content on that instance, it exists on the ActivityPub network." By design, the perimeter of the fediverse is highly permeable; new platforms and users can enter and exit federated systems readily, to both the benefit and detriment of the overall network.

Despite a lack of protocol-mandated governance, many of the more populous parts of the fediverse engage in at least some form of moderation. For example, the Mastodon Server Covenant (which governs whether a Mastodon instance is listed in the central server picker maintained by Mastodon's creator) requires "active moderation against racism, sexism, homophobia and transphobia." While a comprehensive assessment of the policies of federated platforms (including the legitimacy of those policies, and their sufficiency in protecting speech and user safety) is beyond the scope of this article, it is worth noting that where they do exist, the community standards of fediverse instances are often sparse, high-level statements of principle, rather

[1] The terminology used to describe campaigns like the Russian Internet Research Agency (IRA) targeting the 2016 US elections is complex, and increasingly politicized—with terms like "disinformation" now broadly associated with allegations of ideological censorship by technology platforms. Broadly, we use the terms "disinformation," "information operation," "platform manipulation," and "coordinated manipulation" interchangeably throughout this article—though they each refer to slightly different phenomena. More specifically, we draw on a taxonomy of the forms of information disorder originally developed by First Draft, which defines "disinformation" as "content that is intentionally false and designed to cause harm" and "malinformation" as "genuine information that is shared with an intent to cause harm." As we discuss in this article, specific adversarial campaigns like the IRA's efforts may involve a mixture of deceptive behaviors, outright lies, and true information shared to mislead or polarize. In part because of these ambiguities, technology platforms have developed alternate terms—such as Facebook's "coordinated inauthentic behavior"—that characterize such campaigns by the use of deceptive practices like operating multiple social media accounts. As researcher Evelyn Douek has noted, these terms can also be problematic, in large part because of how platform specific they can be, and the difficulties of auditing the standards used by platforms to implement them.

FEDERATED PLATFORMS STATE ASSESSMENT

	CENTRALIZED	FACEBOOK	INSTAGRAM	HORIZON WORLDS	TWITTER	REDDIT	YOUTUBE	DECENTRALIZED	MASTODON	PIXELFED	DIASPORA	PEERTUBE
POLICY												
Public community norms / standards (high level statements)		▶	▶	▶	▶	▶	▶		▶	▶	▶	▶
Public policy explanations with enforcement criteria (detailed)		▶	▶	×	▶	×	▶		×	▷	×	▷
Behavioral manipulation / CIB / platform manipulation policy		▶	▶	×	▶	▷	▶		×	×	×	×
REPORTING												
User reporting capabilities for policy violations		▶	▶	▶	▶	▶	▶		▶	▶	▷	▶
ENFORCEMENT CAPABILITIES												
Permanent account bans		▶	▶	○	▶	▶	▶		▶	▶	▶	▶
Temporary account bans / timeouts		▶	▶	○	▶	▶	▶		▶	▶	×	▶
Ban evasion detection		▷	▷	○	▷	▷	▷		▷	▷	×	▷
Post/content deletion		▶	▶	–	▶	▶	▶		▶	▶	▶	▶
Account visibility restriction		▶	▶	–	▶	▶	▶		▶	×	×	▶
Post/content visibility restriction		▶	▶	–	▶	▶	▶		▶	▶	×	▶
Demonetization		▶	▶	▶	▶	–	▶		–	–	–	–
Automated enforcement tools (heuristics, ML)		▶	▶	×	▶	▶	▶		×	×	×	×
URL blocking		▶	▶	–	▶	▶	▶		×	×	×	×
Media hashing/matching		▶	▶	–	○	▶	▶		×	×	×	×
User-facing moderation controls (block, mute, etc)		▶	▶	▶	▶	▶	▶		▶	▶	×	▶
User identity verification (ID checks, etc)		▶	▶	▶	▶	×	▶		×	×	×	×
Antispam challenges (reCAPTCHA, phone verification)		○	○	–	▶	×	▶		▷	▷	▷	▷
Defederation / instance blocking		–	–	–	–	–	–		▶	○	×	▶
TRANSPARENCY												
Published transparency report		▶	▶	×	▶	▶	▶		×	×	×	×
Terms of service enforcement data		▶	▶	×	▶	▶	▶		×	×	×	×
Behavioral manipulation / CIB / platform manipulation data		▶	▶	×	▷	▶	×		×	×	×	×
Legal information requests data		▶	▶	×	▷	▶	▶		×	×	×	×
Legal removal demands data		▶	▶	×	▷	▶	▶		×	×	×	×
Country/jurisdictional breakdowns of data		▶	▶	×	▷	▶	▶		×	×	×	×
APIs												
Publicly available GET APIs for core platform data (posts, users)		▶	▶	×	▶	▶	▶		▶	▶	▶	▶
Publicly available POST APIs for core platform functions (posts, etc)		▶	▶	×	▶	▶	▶		▶	▶	▶	▶
Publicly available moderation APIs (block, mute, etc)		▶	▶	×	▶	▶	▶		▶	▶	▶	▶

RATING		DESCRIPTION
▶	COMPLETE	The existence of the capability is publicly documented, and is available for use (or has demonstrably/documentably been used) at scale across the platform's core products/business units. *Click on icon for hyperlinked citation.*
▷	PARTIAL	The capability exists, but (1) is not applicable to all of the platform's core products/business units, or (2) has significant functionality gaps that prevent effective use for moderation. *Click on icon for hyperlinked citation.*
×	NONE	The capability does not exist, or exists but cannot be confirmed using publicly available information or expert interviews.
–	N/A	The capability is not applicable/relevant to the platform, based on the properties of the platform (e.g. video moderation for text-only platforms).
○	LIKELY EXISTS	The platform likely possesses the capability, but does not have publicly-listed information on it. *Click on icon for hyperlinked citation.*

than the detailed policies published by larger, centralized platforms. This creates practical ambiguities for the people responsible for moderating content, as well as uncertainty for users about precisely what goes in a given context.

To implement these policies, most federated platforms provide instance administrators and moderators with a rudimentary set of moderation tools. Mastodon, for example, allows moderators to ban individual users, as well as to delete or restrict the visibility of individual pieces of content and accounts.

Federated moderation differs from centralized moderation in an essential way by virtue of the distribution of accounts across multiple instances. A user account has a "local" instance on which it resides, and that instance's moderators have the ability to take direct, destructive action on the user's content (such as deleting it). But, for non-local accounts and content—that is to say, users whose accounts reside on other instances—administrators can only impact their local copies of that content, which influences only the experiences of local users. If a user on instance A encounters a harmful post from a user on instance B, instance A's moderators have no ability to compel instance B to take any action on the harmful post. Mastodon's tools do, however, allow local moderators to take instance-level action to restrict the visibility of content and accounts for all of their local users of an instance, even if that content is permitted on other instances. This has the beneficial effect of giving users greater choice about the policies and governance approaches influencing what they see on social media—but it also makes it more challenging to address fediverse-wide risks created by instances that either cannot or choose not to moderate, whose harmful effects may persist through online and offline action by instance users, even if they are cordoned off from other parts of the fediverse through technical blocks.

In order to address the risks created by specific instances that fail to moderate appropriately, most federated services offer administrators the ability to take moderation action at the instance level, impacting all users on a remote instance, instead of just moderating post by post or account by account. In the case of Mastodon, for example, instances are able to defederate themselves from other servers—in essence, refusing to communicate with or display content from a server deemed to be problematic, rendering its content invisible for all the users on an instance that has chosen to defederate from it. Defederation is largely a server-by-server decision, and, beyond a small number of shared blocklists, few technical or community capabilities exist to deploy these measures at scale across tens of thousands of separate federated instances. Nevertheless, these tactics have, at times, been deployed as a form of broad-based, collective action by fediverse instance moderators—including the notable case of broad defederation from extremist platform Gab.

While defederation offers one scaled mechanism for addressing repeated or prolific harmful conduct, federated platforms largely lack industry-standard capabilities for broad or automated content moderation. Mastodon, for example, does not provide moderators with the ability to block harmful links from being shared on the service. This prevents moderators from being able to ingest lists of known-bad URLs (such as spam and phishing sites) in order to programmatically restrict them. Mastodon also lacks essential tools for addressing media-based harms, like child sexual exploitation, such as media hashing and matching functions (although a number of third parties, including Cloudflare, make such tools available to customers of their content-delivery services). Critically, many of the existing federated platforms have not implemented moderator-facing tools for deploying automation and machine learning to streamline and scale repeated content-moderation actions—functions that are an essential part of the moderation toolkit at all of the existing large, centralized platforms.

As a consequence of the nascent state of moderation capabilities—in terms of both technical features and the manual practices of implementing them—clear governance challenges have emerged for federated service admins. Making moderation a distributed challenge means each instance operator has to reinvent many of the policies and procedures of moderation for themselves. As Ben Werdmuller puts it, "While software is provided to technically moderate, there are very few ecosystem resources to explain how to approach this from a human perspective." The results are predictable for anyone familiar with the challenges of social media content moderation. Users report erroneous or inexplicable bans, with limited recourse from volunteer admins moonlighting as content moderators. Larger-scale harassment campaigns can overwhelm victims and admins alike. Driven by business imperatives, virtually all centralized platforms at least attempt to mitigate these harmful behaviors. But, absent the financial support that goes along with centralized, corporate social media, few parts of the fediverse have been able to successfully marshal the human and technological resources required to successfully execute proactive, accurate content moderation at scale.

THE ABCS OF MODERATING DISINFORMATION

The existing difficulties of moderating federated systems—chief among them, a general lack of resourcing supporting these efforts, even as federated platforms like Mastodon continue to see growth in their adoption—are exacerbated by the highly adaptive nature of coordinated manipulation threats, and the fact that they often require quite different approaches to moderation than those for abuse or hate speech.

To understand these challenges, it's helpful to break down the problem of moderating disinformation into a few components, which researcher Camille François helpfully taxonomized as the "ABCs": actors, behaviors, and content.

CONTENT

Nearly all content-moderation discussions begin with the "C" in François's ABC framework: the content being shared. These analyses focus on the content of a post or account: the language it uses, the links it shares, and the characteristics of a profile. The fundamental challenge of disinformation, however, is that it's seldom apparent from content alone that what you're looking at is actually part of a manipulative campaign. The post- and profile-level evidence most directly available to moderators is rarely dispositive.

Looking back at key examples of the Russian Internet Research Agency (IRA) activity on Twitter in 2016, what is most striking about posts from prominent accounts like "Crystal Johnson," a Russian persona purporting to be an African-American woman, is that, by and large, the content of the posts was true: while the IRA's earliest efforts involved comically ineffective rumormongering about an alleged outbreak of Ebola in Atlanta, the bulk of its activity during and after the 2016 US elections used a more subtle tactic of sharing factually accurate but divisive rhetoric using inauthentic behaviors (such as fake accounts). This stymied efforts to moderate content based on policies that evaluate the substance of a post.

Even when we know content has been created by a troll farm, addressing it as content is challenging (if not impossible). For example, researcher Josh Russell captured hundreds of examples of memes created by the IRA on Instagram in 2018; those same exact memes resurfaced a year later in a network of spammy Facebook pages operated out of Ukraine. If Meta, possessing all the relevant data about these campaigns and having extensive capabilities to detect similar media, couldn't catch this, how can we expect Mastodon instance moderators to keep pace, particularly given the lack of media-hashing and matching functions? And even if these capabilities were developed and implemented, matching-based approaches are inherently reactive to already-known examples of disinformation; modifications to existing assets, or the creation of novel content, would quickly undermine the effectiveness of these approaches.

These challenges are exacerbated by the phenomenon of real people authentically sharing content from, or similar to, trolls. Many of the memes originally created by IRA staff in St. Petersburg continue to circulate on social media among folks who just happen to think they're funny or interesting. If real people intentionally choose to amplify messaging sourced from, or consistent with, a government-sponsored trolling campaign, it's not obvious what, if anything, moderators should do. While some have argued that stronger media literacy would greatly help with this issue, the particular characteristics of IRA-style inauthenticity makes this a challenging proposition. What forms of literacy would help most people recognize a Crystal Johnson-style account as inauthentic using publicly available data? The sparseness of profiles on services like Mastodon, as well as their relative newness, makes it harder to make genuinely informed judgments—limitations that apply to users and moderators alike.

BEHAVIOR

At the core of disinformation campaigns is the concept of manipulative behavior: the practice of engaging in tactics of sharing and disseminating content that seek to inauthentically propagate, promote, or inflate the reach of an account or piece of content. As François puts it in the ABC framework: "At the end of the day, deceptive behaviors have a clear goal: to enable a small number of actors to have perceived impact that a greater number of actors would have if the campaign were organic." Put another way, a few staffers at a troll farm in St. Petersburg are unlikely to be particularly influential absent behavior that skews the attention economy of social media in their favor.

In many cases, this is just another way of referring to spam.[2] High-volume, low-sophistication political-manipulation campaigns became a feature of Twitter in particular—with threat actors based in Venezuela and China (to give just two examples) deploying them with some regularity.

Federated services, at least in their present implementations, have some inherent resilience to these tactics. The lack of algorithmic recommendations means there's less of an attack surface for inauthentic engagement and behavioral manipulation. While Mastodon has introduced a version of a "trending topics" list—the true battlefield of Twitter manipulation campaigns, where individual posts and behaviors are aggregated into a prominent, platform-wide driver of attention—such features tend to rely on aggregation of local (rather than global or federated) activity, which removes much of the incentive for engaging in large-scale spam. There's not really a point to trying to juice the metrics on a Mastodon post or spam a hashtag, because there's no algorithmic reward of attention for doing so. The lack of built-in monetization programs on virtually all federated platforms—at least presently—likewise reduces incentives for programmatic malfeasance.

These disincentives for manipulation have their limits, though. Some of the most successful disinformation campaigns on social media, like the IRA's use of fake accounts, relied less on spam and more on the careful curation of individual "high-value" accounts—with uptake of their content being driven by organic sharing, rather than algorithmic amplification. Disinformation is just as much a community problem as it is a technological one (i.e., people share content they're interested in or get emotionally activated by, which sometimes originates from troll farms)—which can't be mitigated just by eliminating the algorithmic drivers of virality.

Detection of behavioral manipulation relies, in large part, on access to data about on-platform activity—and the openness of federated platforms has largely resulted in the ready availability of application programming interfaces (APIs) to enable this kind of access. For example, Mastodon has a robust set of public APIs that would allow researchers to study the conversations happening on the service. But federation complicates the use of these APIs to study ecosystem-level threats. Whereas Twitter's APIs offer a single channel for col-

[2] In the early days of Congress investigating Russian interference in the 2016 US election, Twitter staff briefed stakeholders on Capitol Hill about the company's efforts to combat what we were calling "political spam." We were excoriated by a few of the people with whom we spoke, who said that even calling it "spam" meant we were missing the gravity of the situation. Twitter subsequently came up with the term "platform manipulation" as an alternative that would signal how seriously we took the issue. See: Patrick Conlon, William Nuland, and Kanishk Karan, "Investigating Influence Operations By Twitter Integrity," in Victoria Smith, Jon Bateman, and Dean Jackson, eds., *Perspectives for Influence Operations Investigators* (Washington DC: Carnegie Endowment for International Peace, 2022).

lecting data about all the activity happening globally across the Twitter service, Mastodon's APIs are mostly instance specific. As a result, many data-collection efforts either involve focusing on a handful of the largest instances, or needing to go down an essentially limitless rabbit hole of collecting data from successively smaller and smaller instances until you reach a point of diminishing returns—with no guarantee that the threats you're hunting aren't lurking on the *n+1*th instance from which you'd collect data.

The federated nature of this threat creates similar challenges for moderators. Many of the techniques employed by large platforms to detect manipulation involve surveying the full population of accounts and activity, and looking for unusual clusters or patterns of behavior within that population—a practice of threat identification using centralized telemetry. To give a rudimentary example: if you look at posts containing a hashtag like #ElectionNight2022, group those posts by the Internet Protocol (IP) address from which they were sent, and observe that a bunch of them were sent from an IP address in Russia, you might investigate the accounts responsible to see if something fishy is going on. But in a federated system, instance admins only have comprehensive logs for the activity of local users of their particular instance—which means a threat actor who spreads their inauthentic accounts across a handful of the biggest instances is both less likely to be caught as behaviorally anomalous and less likely to have the full scope of their operation, across all the instances on which they operate, be detected. An analyst is less likely to spot a suspicious cluster of accounts if it's just one or two users among tens of thousands.

This also assumes that instance moderators have the time, knowledge, tools, and governance frameworks necessary to do the highly specialized work of disinformation detection and analysis. Training programs at large platforms to get even technically proficient analysts fully up to speed on advanced analytic techniques can take months. There are also costs beyond just time and attention. Even if you assume that moderators have the necessary technical skills to do this work, the compute costs alone of querying against these large-scale datasets are considerable; we can't reasonably expect volunteers to do this work pro bono. These capabilities beget fraught privacy and user-control challenges as well. What safeguards exist to ensure instance admins, or their designees, engage only in appropriate uses of sensitive user logs?

Finally, there's also the challenge of how, exactly, to moderate a distributed but coordinated threat. Mastodon's moderation capabilities provide for a few rudimentary anti-spam techniques for addressing scaled threats (techniques even the Mastodon documentation notes will be circumvented by dedicated spammers)—but Mastodon moderation is focused largely on either dealing with individually problematic users (by restricting or banning them from a given instance) or the radical option of defederating a wholly problematic instance. Spam and platform manipulation are unlikely to be solvable using this tactic, because they primarily manifest as distributed threats across mainstream, non-malicious instances. Put another way, we shouldn't expect sophisticated adversarial threats to concentrate themselves on single instances, waiting to be defederated. Instead, inauthentic accounts are likely to be dispersed across mainstream servers. This creates a distributed burden of detection across already-overworked and under-equipped moderators, who need to deal with these accounts one by one, instance by instance.

This is a social challenge as well as a technical one. The ActivityPub protocol gives instance operators ways to defend themselves against bad actors by defederating problematic servers, but what is the appropriate course of action when well-intentioned admins may be unaware of, or unable to meaningfully address, the malicious activity they host? As with the challenges of addressing the attempted infiltration of social movements by troll farms, it isn't always clear how to know which instances and individuals are trustworthy—a dynamic that malign actors can exploit, as happened with a Ukrainian Mastodon instance following the Russian invasion of Ukraine. At their worst, these dynamics may lead to otherwise legitimate instances being defederated or restricted for failing to appropriately moderate—to the detriment of their other legitimate users.

Among the most commonly proposed solutions to these issues is making Mastodon instances invite only, or requiring some kind of trusted referral model for new signups. This may well be a viable solution for parts of the fediverse that intentionally prioritize small community size and affinity based on identity or interest. But the “gated community” model has at least three key challenges as a broader strategy. First, this only solves the problem of “local” manipulation, not the impacts of federated behavior on non-local viewers of that content. Second, it’s not clear that this is actually a way to address the most sophisticated and insidious forms of manipulative behavior. Elaborately constructed inauthentic profiles—like Crystal Johnson, or the deep, cross-platform persona development tactics described in a recent expose about an Israeli disinformation purveyor—will often withstand anything but the most invasive forms of validation. (And, inevitably, the more invasive validation becomes, the less usable a service is by vulnerable people and groups, who might have good reasons for not wanting to disclose their personal information to instance operators they don’t know or trust.) Finally, and most fundamentally, for people looking to Mastodon and the fediverse as an alternative to centralized social platforms like Twitter, raising barriers to entry introduces fundamental tradeoffs against the very network effects that could help make Mastodon a mass-market product.

ACTORS

In François’s framework, “actors” refers to the people or groups behind deceptive activity. The basic premise is that it matters who or what is engaged in malicious or harmful conduct. Often, actor-level analysis is reduced to the security practice of “attribution”—the name-and-shame exposure of the individuals or groups responsible for an attack or intrusion. Certainly, attribution has an important part to play in counteracting disinformation; it gives nation-states critical evidence needed to enact offline consequences for online malfeasance. We don’t need to look further than the dozens of indictments of Russian operatives following the 2016 elections to see this direct interplay of platform-based investigations and offline law-enforcement action.

But attribution is far from the only goal of actor-level analysis. Understanding the actors responsible for a disinformation campaign meaningfully influences how platforms respond—and can give platforms necessary tools for addressing these challenges in a scalable way. The present state of fediverse moderation—from both labor and technological perspectives—has two primary structural constraints on actor-level analysis: a lack of capability and capacity for longitudinal enforcement; and a lack of collaboration with other groups tracking the same threat actors, which results in inefficiency and detection gaps.

Longitudinal analysis refers to tracking and analyzing the behavior of specific threat actors or patterns of malicious behavior over time. These practices have typically been carried out by platforms themselves, and by a wide array of civil-society and academic groups. These are not abstract pursuits; they have specific, practical applications that meaningfully contribute to platform capabilities to mitigate disinformation. Principally, understanding the behaviors and motivations of persistent threats helps platforms develop effective mitigation strategies suited to applying optimal, cost-effective pressure to a particular actor based on their unique goals and constraints. For example, cost-optimizing threat actors—like many of the commercially motivated groups peddling political spam in order to sell t-shirts or redirect traffic to ad-filled content farms—will be most impacted by enforcement strategies that raise the cost of doing business. The individual unit costs of spam are low, but so, too, are the gains realized through these campaigns. Strategically imposing additional expenses through mechanisms like mandatory phone-number verification, completion of CAPTCHA challenges, and domain blocking can, over time, make it cost-ineffective for financial spam campaigns to target an effective platform. But these enforcement measures have user-experience and privacy tradeoffs, and platforms generally avoid applying them indiscriminately for fear of alienating users—which requires a targeted approach that seeks out and enforces against specific threats in specific ways.

The current state of fediverse moderation has two key constraints on the ability to enact this kind of targeted pressure on adversarial behavior. First, actor-level analysis requires time-consuming and labor-intensive tracking and documentation. Differentiating between a commercially motivated spammer and a state-backed troll farm often requires extensive research, extending far beyond activity on one platform or website. The already unsustainable economics of fediverse moderation seem unlikely to be able to accommodate this kind of specialized investigation.

Second, even if you assume moderators can, and do, find accounts engaged in this type of manipulation—and understand their actions and motivations with sufficient granularity to target their activity—the burden of continually monitoring them is overwhelming. Perhaps more than anything else, disinformation campaigns demonstrate the “persistent” in “advanced persistent threat”: a single disinformation campaign, like China-based Spamouflage Dragon, can be responsible for tens or even hundreds of thousands of fake accounts per month, flooding the zone with low-quality content. The moderation tools built into platforms like Mastodon do not offer appropriate targeting mechanisms or remediations to moderators that could help them keep pace with this volume of activity. Moderation actions are wholly manual, and are limited to either banning or restricting individual accounts, or blocking entire ranges of IP addresses or email domains. Moderators lack the capability to deploy heuristics—essentially, sets of rules that describe patterns of adversarial behavior—that can automate these actions. Without these capabilities to automate enforcement based on long-term adversarial understanding, the unit economics of manipulation are skewed firmly in favor of bad actors, not defenders.

These are solvable product and engineering challenges—and no doubt the moderation tools built into Mastodon and other federated products will improve over time. But there are critical labor components as well. In the case of a persistent but poorly obfuscated campaign like Spamouflage Dragon, detection isn’t especially difficult. But as the tactics and thematic focus of the campaign evolves over time, scaled remediation requires continual maintenance to keep things from going off the rails. Heuristics that are viable one day can become inaccurate the next. Machine-learning models exhibit drift over time and can either under-detect or over-detect the target activity. “Set it and forget it” is not a viable strategy for dealing with dedicated adversaries. Look no further than Twitter following the dismissal of staff responsible for monitoring Chinese-language disinformation, heuristics developed to address Spamouflage Dragon and other campaigns have, according to reports, declined in accuracy to a point where the legitimate accounts of activists and users are being inaccurately restricted or banned. Responsible deployment of even sophisticated technical enforcement capabilities requires ongoing, sustained effort by moderators.

Critically, efforts to disrupt persistent threat actors are most successful when approached at a community or ecosystem level, rather than by individual platforms in isolation. Many of the most prolific disinformation campaigns are notable for their presence across multiple platforms. Russian campaigns targeting the White Helmets in Syria in 2017 and 2018 spanned Twitter, YouTube, and mainstream and alternative media properties. More recently, the Secondary Infektion operations promoting Russian interests spanned more than three hundred sites and platforms. From an analytic perspective, it can be challenging, if not impossible, to recognize individual accounts or posts as connected to a disinformation campaign in the absence of cross-platform awareness of related conduct. The largest platforms—chiefly, Meta, Google, and Twitter (pre-acquisition)—regularly shared information, including specific indicators of compromise tied to particular campaigns, with other companies in the ecosystem in furtherance of collective security. Information sharing among platform teams represents a critical way to build this awareness—and to take advantage of gaps in adversaries’ operational security to detect additional deceptive accounts and campaigns.

Federated moderation makes this kind of cross-platform collaboration difficult. Thousands of individual instance operators each have responsibility for a potential target of this conduct, but it’s infeasible for larger platforms, like Meta and Google, to engage with moderators or admins from each instance directly. Even assuming these engagements are limited to a handful of the largest fediverse instances, the legal frameworks and contractual protections needed to share data across platforms without running afoul of privacy regulations like the General Data Protection Regulation (GDPR) require specialized legal expertise and negotiation, which are often out of reach for hobbyist efforts. In addition, absent an institutionalized way to verify the trustworthiness and legitimacy of instance admins and moderators, larger platforms will have limited information on who they are working with—and, correspondingly, may either choose not to engage or feel constrained in their ability to share relevant data. Bad actors posing as moderators of legitimate fediverse instances can leverage these structural ambiguities to gain access to larger platforms' staff and intel, creating commercial, political, and privacy risk.

Even among federated services, it's challenging for instance moderators to engage with each other in a structured way to counteract shared threats. Collaborative security models are common both within the social media industry and outside of it—including financial-intelligence units in the financial-services sector, and information-sharing and analysis centers in the information security context. These institutionalized collaborations are predicated on a high degree of alignment about the scope and nature of the threats in question. While decentralized community governance has had notable successes on platforms like Wikipedia, the notable lack of agreement on norms and standards across instances makes it challenging for these collaborative practices to adapt to the fediverse. For example, on Mastodon, tagging discussions using "#fediblock" has emerged as a grassroots practice for sharing information about bad actors, but these approaches have run up against the challenges of a fully distributed, low-trust model. Moderators report that it's hard to know which accounts of bad behavior are trustworthy or verified enough to warrant enforcement without firsthand confirmation.

This is a product not only of technical capabilities, but of the cultural norms of federated systems, exposing a core challenge in establishing effective collaboration. Evelyn Douek has written critically about the so-called "content cartels" that form when mainstream platforms collaborate with each other; federated approaches, in part, offer an alternative configuration. When platforms are designed and built to empower individuals and communities to be self-sovereign, as in the case of many federated services, their operators and moderators may be reasonably skeptical of the kind of centralized designations inherent in this kind of information sharing. It would hardly be desirable for instance operators to uncritically import enforcement decisions and bad-actor designations without ensuring that these data are both aligned with instance policies, and are sourced from trustworthy moderators. A failure state for the promise of the fediverse is homogeneity of moderation as a product of convenience. But leaving it to individual moderators to assess, designate, and track troll farms and other bad actors for themselves is hardly a reasonable alternative.

Establishing mechanisms for transparency reporting in the fediverse could help address difficulties in moderating across instances. This may soon become necessary, as the European Digital Services Act is likely to classify Mastodon instances as independent "online platforms" subject to transparency reporting obligations. No such reporting practice currently exists on major fediverse platforms, and the creation thereof would only be complicated by the need for compliance and coordination across moderators and admins, and the lack of a centralized structure to report on this information.

NEXT STEPS

As consumers explore alternatives to mainstream social media platforms, malign actors will migrate along with them—a form of cross-platform regulatory arbitrage that seeks to find and exploit weak links in our collective information ecosystem. Further research and capability building are necessary to avoid the further proliferation of these threats.

A critical element of this work is identifying the specific intervention strategies that are suited to the sociotechnical properties of federated and distributed social platforms. Key research questions include the following.

- What are the risks and challenges posed by disinformation and manipulative behavior on federated platforms, and how do these risks differ from those created on centralized social media services?

- What policies and governance approaches currently exist for manipulative behavior and disinformation across major instances of federated services?
- How do the users and operators of federated services conceptualize the risks of manipulative conduct, given the norms and governance structures of existing decentralized communities?
- What are the existing moderation capabilities built into federated services, and how effective are they at addressing behavioral and scaled threats?
- What technical capabilities—moderation tools, datasets, APIs, etc.—are required to effectively manage coordinated manipulation and disinformation threats?
- What moderation and analytic capabilities are necessary to help instance operators, moderators, and the users of federated services address the risks and threats created by persistent adversarial behavior?
- What are the appropriate governance frameworks and organizational structures for this work in a decentralized context?

Answers to these questions will help structure responses across three critical constituencies: the developers of open-source fediverse services, and the developers of complementary tools and features that enable effective moderation of federated social media; the individuals and groups engaged in investigations, analysis, and moderation of federated services; and investors, funders, and donors engaged with platform governance and counter-manipulation efforts.

AUTHORSHIP AND ACKNOWLEDGEMENTS

This report was authored by Yoel Roth and Samantha Lai. Yoel Roth is a technology policy fellow at the University of California, Berkeley, and a nonresident scholar at the Carnegie Endowment for International Peace. He was previously the head of trust and safety at Twitter, and received his PhD from the Annenberg School for Communication at the University of Pennsylvania. Samantha Lai is a research analyst with the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace. Prior to joining Carnegie, she was a research analyst at the Brookings Institution and the Foreign Policy Research Institute. This work was supported by the William and Flora Hewlett Foundation, the John S. and James L. Knight Foundation, and the Carnegie Endowment for International Peace. The authors are grateful to Patrick Conlon, Renee DiResta, Camille François, Jeff Jarvis, Jaz-Michael King, Hilary Ross, and members of the Atlantic Council Task Force for a Trustworthy Future Web for their feedback and perspectives.