MODERATING THE FEDIVERSE: CONTENT MODERATION ON DISTRIBUTED SOCIAL MEDIA

Alan Z. Rozenshtein[*]

INTRODUCTION

Current approaches to content moderation generally assume the continued dominance of "walled gardens": social-media platforms that control who can use their services and how. Whether the discussion is about self-regulation, quasi-public regulation (e.g., Facebook's Oversight Board), government regulation, tort law (including changes to Section 230), or antitrust enforcement, the assumption is that the future of social media will remain a matter of incrementally reforming a small

[*] Associate Professor of Law, University of Minnesota. For helpful comments I thank Laura Edelson, Kyle Langvardt, Erin Miller, Chinmayi Sharma, and participants at the Big Tech and Antitrust Conference at Seton Hall Law School, the Information Society Project and the Freedom of Expression Scholars Conference at Yale Law School, the Association for Computing Machinery (ACM) Symposium on Computer Science and Law, and the Max Weber Programme Multidisciplinary Research Workshop at the European University Institute. For excellent research assistance I thank Caleb Johnson and Isabel Park.

This Essay will be republished as a book chapter in MEDIA AND SOCIETY AFTER TECHNOLOGICAL DISRUPTION (Gus Hurwitz & Kyle Langvardt eds., forthcoming Cambridge Univ. Press 2023).

group of giant, closed platforms. But, viewed from the perspective of the broader history of the Internet, the dominance of closed platforms is an aberration. The Internet initially grew around a set of open, decentralized applications, many of which remain central to its functioning today.

Email is an instructive example. Although email is hardly without its content-moderation issues—spam, in particular, has been an ongoing problem—there is far less discussion about email's content-moderation issues than about social media's. Part of this is because email lacks some of the social features that can make social media particularly toxic. But it is also because email's architecture simply doesn't permit the degree of centralized, top-down moderation that social-media platforms can perform. If "ought" implies "can," then "can't" implies "need not." There is a limit to how heated the debates around email-content moderation can be, because there's an architectural limit to how much email moderation is possible. This raises the intriguing possibility of what social media, and its accompanying content-moderation issues, would look like if it too operated as a decentralized protocol.

Fortunately, we don't have to speculate, because decentralized social media already exists in the form of the "Fediverse"—a portmanteau of "federation" and "universe." Much like the decentralized infrastructure of the Internet, in which the HTTP communication protocol facilitates the retrieval and interaction of webpages that are stored on servers around the world, Fediverse protocols power "instances," which are comparable to social-media applications and services. The most important Fediverse protocol is ActivityPub, which powers the most popular Fediverse apps, notably the Twitter-like microblogging service Mastodon, which has over a million active users and continues to grow, especially in the wake of Elon Musk's purchase of Twitter.[1]

The importance of decentralization and open protocols is increasingly recognized within Silicon Valley. Twitter co-founder Jack Dorsey has launched Bluesky, a Twitter competitor built on the decentralized ATProtocol. Meta's Mark Zuckerberg has described his plans for an "open, interoperable metaverse" (though how

[1] *See* Barbara Ortutay, *Twitter Drama Too Much? Mastodon, Others Emerge as Options*, AP NEWS (Nov. 12, 2012), https://perma.cc/PY4F-8GD9.

far this commitment to openness will go remains to be seen).[2] And established social media platforms are building in interoperability with ActivityPub applications.[3]

Building on an emerging literature around decentralized social media,[4] this brief essay seeks to give an overview of the Fediverse, its benefits and drawbacks, and how government action can influence and encourage its development. Part I describes the Fediverse and how it works, first distinguishing open from closed protocols and then describing the current Fediverse ecosystem. Part II looks at the specific issue of content moderation on the Fediverse, using Mastodon as a case study to draw out the advantages and disadvantages of the federated content-moderation approach as compared to the currently dominant closed-platform model. Part III considers how policymakers can encourage the Fediverse through participation, regulation, antitrust enforcement, and liability shields.

I. CLOSED PLATFORMS AND DECENTRALIZED ALTERNATIVES

A. A Brief History of the Internet

A core architectural building block of the Internet is the *open protocol*. A *protocol* is a rule that governs the transmission of data. The Internet consists of many such protocols, ranging from those that direct how data is physically transmitted to those that govern the most common Internet applications, like email or web browsing. Crucially, all these protocols are *open*, in that anyone can set up and operate a router, website, or email server without needing to register with or get permission

---

[2] Andrew Hayward, *An 'Open, Interoperable' Metaverse is 'Better for Everyone': Meta's Mark Zuckerberg*, YAHOO! NEWS (Oct. 11, 2022), https://perma.cc/E32U-FQ7C.

[3] David Pierce, *Can ActivityPub Save the Internet?*, VERGE (Apr. 20, 2023).

[4] *See, e.g.*, Mike Masnick, *Protocols, Not Platforms: A Technological Approach to Free Speech*, KNIGHT FIRST AMEND. INST. (Aug. 21, 2019), https://perma.cc/J2QD-YVF7; FRANCIS FUKUYAMA ET AL., STANFORD CYBER POL'Y CTR., MIDDLEWARE FOR DOMINANT DIGITAL PLATFORMS: A TECHNOLOGICAL SOLUTION TO A THREAT TO DEMOCRACY (2021), https://perma.cc/S54K-JVEX; Daphne Keller, *The Future of Platform Power: Making Middleware Work*, 32 J. DEMOCRACY 168 (2021); Chand Rajendra-Nicolucci & Ethan Zuckerman, *What If Social Media Worked More Like Email?*, *in* AN ILLUSTRATED FIELD GUIDE TO SOCIAL MEDIA 24 (Chand Rajendra-Nicolucci & Ethan Zuckerman eds., 2021), https://perma.cc/F3LC-LGR4; Robert W. Gehl & Diana Zulli, *The Digital Covenant: Non-Centralized Platform Governance on the Mastodon Social Network*, INFO., COMMC'N & SOC'Y (forthcoming), https://perma.cc/H4XN-9E9K.

from a central authority.[5] Open protocols were key to the first phase of the Internet's growth because they enabled unfettered access, removing barriers and bridging gaps between different communities. This enabled and encouraged interactions between groups with various interests and knowledge, resulting in immense creativity and idea-sharing.

But starting in the mid-2000s, a new generation of closed platforms—first Facebook, YouTube, and Twitter, and later Instagram, WhatsApp, and TikTok—came to dominate the Internet habits of most users.[6] Today's Internet users spend an average of seven hours online a day, and approximately 35% of that time is spent on closed social-media platforms.[7] Although social-media platforms use the standard Internet protocols to communicate with their users—from the perspective of the broader Internet, they just operate as massive web servers—their internal protocols are closed. There's no Facebook protocol that you could use to run your own Facebook server and communicate with other Facebook users without Facebook's permission. Thus, major social-media platforms are the most important example of the Internet's steady, two-decades-long takeover by "walled gardens."[8]

There are many benefits to walled gardens; otherwise, they wouldn't have taken over. Closed systems are attractive for the companies that run them because the companies can exert greater control over their platforms through content and user moderation. But the draw for platform owners is insufficient; only by providing users with a better experience (or at least convincing them that their experience is better) could closed platforms have come to dominate social media.

---

[5] The distinction between open and closed protocols is not clear-cut. Some of the core technology behind the Internet—for example, the Domain Name System, which maps IP addresses to human-readable domain names—has a centralized registration system. But this system imposes relatively minimal control, and the entity that runs it, the Internet Corporation for Assigned Names and Numbers (ICANN), is a multistakeholder nonprofit that prioritizes openness and interoperability.

[6] An early challenge to the open Internet came from the first generation of giant online services providers like America Online, Compuserve, and Prodigy, which combined dial-up Internet access with an all-encompassing web portal that provided both Internet content and messaging. But as Internet speeds increased and web browsing improved, users discovered that the limits of these closed systems outweighed their benefits, and they faded into irrelevance by the 2000s.

[7] Simon Kemp, *Digital 2022: Global Overview Report*, DATAREPORTAL (Jan. 26, 2022), https://perma.cc/XM4G-DLND.

[8] The other major example of a move to closed system is the dominance of smartphones, which (especially iOS devices) are far more closed than are personal computers.

Closed platforms have indeed often provided more value to users. The logic of enclosure applies as much to virtual spaces as it does to real ones: Because companies can more thoroughly monetize closed platforms, they have a greater incentive to invest more in those platforms and provide better user experiences. One can create a Twitter account and begin posting tweets and interacting with others within minutes; good luck setting up your own microblogging service from the ground up. And because companies have full control over the platform, they can make changes more easily—thus, at least in the short term, closed platforms can improve at a faster rate than can open platforms, which often struggle with cumbersome, decentralized consensus governance.

Most important, at least from the perspective of this essay, is closed platforms' advantages when it comes to moderation. Closed platforms can be moderated centrally, which enables greater control over what appears on the network. And the business models of closed platforms allow them to deploy economic and technological resources at a scale that open, decentralized systems simply cannot match. For example, Meta, Facebook's parent company, has spent over $13 billion on "safety and security" efforts since the 2016 election, employing, both internally and through contractors, 40,000 employees on just this issue. And Meta's investments in AI-based content-moderation tools have led it to block billions of fake accounts.[9] Content moderation, as Tarleton Gillespie notes, "is central to what platforms do, not peripheral" and "is, in many ways, the commodity that platforms offer."[10] Indeed, this concern with security—whether about malicious code, online abuse, or offensive speech—is one of the most important drivers of the popularity of closed systems.[11]

But closed platforms have become a victim of their own success. They have exacerbated the costs of malicious action by creating systems that are designed to be as frictionless as possible within the network (even if access to the network is controlled by the platform). At the same time, they have massively increased user expectations regarding the moderation of harmful content, since centralization allows

---

[9] *Our Progress Addressing Challenges and Innovating Responsibly*, FACEBOOK (Sept. 21, 2021), https://perma.cc/3FHT-3TB8.

[10] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 13 (2018).

[11] JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET AND HOW TO STOP IT 59 (paperback ed. 2008).

(in theory, though not in practice) the complete elimination of harmful content in a way that the architecture of an open system does not. Closed platforms impose uniform, top-down standards, which inevitably leave many users unsatisfied. And they raise concerns about the handful of giant companies and Silicon Valley CEOs exercising outsized control over the public sphere.[12]

In other words, large, closed platforms are faced with what might be called the *moderator's trilemma*. The first prong is that platform userbases are large and diverse. The second prong is that the platforms use centralized, top-down moderation policies and practices. The third prong is that the platforms would like to avoid angering large swaths of their users (not to mention the politicians that represent them). But the content-moderation controversies of the past decade suggest that these three goals can't all be met. The large closed platforms are unwilling to shrink their user bases or give up control over content moderation, so they have tacitly accepted high levels of dissatisfaction with their moderation decisions. The Fediverse, by contrast, responds to the moderator's trilemma by giving up on centralized moderation.

B. The Fediverse and Its Applications

The term "Fediverse" refers collectively to the protocols, servers, applications, and communities that enable decentralized social media. The most popular of these protocols is ActivityPub, which is developed by the World Wide Web Consortium, the main international standards organization for the World Wide Web, and which has also developed the HTML, XML, and other foundational Internet standards.[13]

To understand how ActivityPub operates, it's important to appreciate that all

---

[12] When Elon Musk first made his bid to purchase Twitter, Twitter co-founder Jack Dorsey tweeted:

> In principle, I don't believe anyone should own or run Twitter. It wants to be a public good at a protocol level, not a company. Solving for the problem of it being a company however, Elon is the singular solution I trust. I trust his mission to extend the light of consciousness.

@jack, TWITTER (Apr. 25, 2022, 9:03 PM), https://perma.cc/VD56-QNRQ. The chaos that has roiled Twitter since Musk's takeover suggests that Dorsey's faith in Musk's "mission to extend the light of consciousness" was misplaced while underscoring the observation that Twitter would be better as "a public good at a protocol level, not a company." To his credit, Dorsey has since recognized Musk's faults as Twitter's owner. *See* Faiz Siddiqui & Will Oremus, *Twitter Founder Jack Dorsey Says Musk Wasn't an Ideal Leader After All*, WASH. POST (Apr. 29, 2023).

[13] *ActivityPub*, W3C, https://perma.cc/L84U-C5D6.

social-media platforms are built around the same core components: users creating and interacting with pieces of content, whether posts (Facebook), tweets (Twitter), messages (WhatsApp), images (Instagram), or videos (YouTube and TikTok). When a user tweets (for example), they first send the tweet to a Twitter server. That Twitter server then distributes that tweet through the Twitter network to other users. Like all platforms, Twitter has its own internal protocol that processes the data representing the tweet: the tweet's content plus metadata like the user handle, the time the tweet was made, responses to the tweet ("likes" and "retweets"), and any restrictions on who can see or reply to the tweet.

ActivityPub generalizes this system. The ActivityPub protocol is flexible enough to accommodate different kinds of social-media content. This means that developers can build different applications on top of the single ActivityPub protocol; thus, Friendica replicates the main features of Facebook, Mastodon replicates those of Twitter, and PeerTube of YouTube. But unlike legacy social-media platforms, which do not naturally interoperate—one can embed a YouTube link in a tweet, but Twitter sees the YouTube content as just another URL, rather than a type of content that Twitter can directly interact with—all applications built on top of ActivityPub have, in principle, access to the same ActivityPub data, allowing for a greater integration of content.[14]

The most important feature of ActivityPub is that it is decentralized. The servers that users communicate with and that send content around the network are independently owned and operated. Anyone can set up and run an ActivityPub server—generally called an "instance"—as long as they follow the ActivityPub protocol. This is the key feature distinguishing closed platforms like Twitter or Facebook from open platforms like ActivityPub—or email or the World Wide Web, for that matter: Anyone can run an email or web server if they follow the relevant protocols.

ActivityPub's decentralized nature means that each instance can choose what content flows across its network and use different content-moderation standards. An instance can even choose to block certain users, types of media (e.g., videos or

---

[14] For example, as PeerTube, a video-sharing platform, notes, "you can follow a PeerTube user from Mastodon (the latest videos from the PeerTube account you follow will appear in your feed), and even comment on a PeerTube-hosted video directly from your Mastodon's account." PEERTUBE, https://perma.cc/RT9C-9TVH.

images), or entire other instances. At the same time, each instance's content-moderation decisions are locally scoped: 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.

This leads to a model of what I call *content-moderation subsidiarity*. Just as the general principle of political subsidiarity holds that decisions should be made at the lowest organizational level capable of making such decisions,[15] content-moderation subsidiarity devolves decisions to the individual instances that make up the overall network.

A key guarantor of content-moderation subsidiarity is the ability of users to switch instances if, for example, they are dissatisfied with how their current instance moderates content. If a user decides to move instances, their followers will automatically refollow them at their new account.[16] Thus, migrating from one Mastodon instance to another does not require starting from scratch. The result is that, although Fediverse instances show some of the clustering that is characteristic of the Internet as a whole,[17] no single instance monopolizes the network.[18]

Using Albert Hirschman's theory of how individuals respond to dissatisfaction with their organizations,[19] we can say that the Fediverse empowers users to exercise powers of *voice* and *exit* more readily and meaningfully than they could on a centralized social-media platform. Rather than simply put up with dissatisfactions, the

---

[15] *See generally* Andreas Føllesdal, *Subsidiarity*, 6 J. POL. PHIL. 190 (1998).

[16] Mastodon does not currently allow moving posts from one instance to another, but it does allow users to download a record of their posts. *How to Migrate from One Server to Another*, MASTODON, https://perma.cc/Y4XY-KM6W.

[17] *See* Lada A. Adamic & Bernardo A. Huberman, *Zipf's Law and the Internet*, 3 GLOTTOMETRICS 143, 147–48 (2002), https://perma.cc/H8LL-G9LY ("[T]here are many small elements contained within the Web, but few large ones. A few sites consist of millions of pages, but millions of sites only contain a handful of pages. Few sites contain millions of links, but many sites have one or two. Millions of users flock to a few select sites, giving little attention to millions of others.").

[18] A list of Mastodon instances, sorted by number of users, is available/^e at https://perma.cc/S8JU-GGTW.

[19] *See* ALBERT O. HIRSCHMAN, EXIT, VOICE, AND LOYALTY: RESPONSES TO DECLINE IN FIRMS, ORGANIZATIONS, AND STATES (1970).

Fediverse permits users to choose the instance that best suits them (exit) and to use that leverage to participate in instance governance (voice). Of course, users on closed platforms can (and frequently do) express their grievances with how the platform is moderated—perhaps most notably on Twitter, where a common (and ironic) subject for tweets is how terrible Twitter is—but such an "affective voice" is far less likely to lead to meaningful change than the "effective voice" that the Fediverse enables.[20]

Some existing companies, though they remain centralized in most respects, have enhanced users' voice and exit privileges by decentralizing their platform's moderation practices. For example, Reddit, the popular message-board platform, grants substantial autonomy to its various subreddits, each of which has its own moderators. Indeed, Reddit is frequently held up as the most prominent example of bottom-up, community-based content moderation.[21] One might thus ask: does the Fediverse offers anything beyond what already exists on Reddit and other sites, like Wikipedia, that enable user-led moderation?

Indeed it does, because the Fediverse's decentralization is a matter of *architecture*, not just policy. A subreddit moderator has control only insofar as Reddit, a soon-to-be public company,[22] permits that control. Because Reddit can moderate any piece of content—and can even ban a subreddit outright—no matter whether the subreddit moderator agrees, the company is subject to public pressure to do so. Perhaps the most famous example is Reddit's banning of the controversial pro-Trump r/The_Donald subreddit several months before the 2020 election.[23]

Taken as a whole, the architecture of the Fediverse represents a challenge not only to the daily operations of incumbent platforms, but also to their very theoretical bases. Media scholars Aymeric Mansoux and Roel Roscam Abbing have developed what is so far the most theoretically sophisticated treatment of the Fediverse's content-moderation subsidiarity, which they characterize as a kind of "agonism":

---

[20] *See* Seth Frey & Nathan Schneider, *Effective Voice: Beyond Exit and Affect in Online Communities*, NEW MEDIA & SOC'Y (Sept. 2021), https://perma.cc/VQ6K-6CBY.

[21] *See, e.g.*, James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42, 94–101 (2015).

[22] Cory Weinberg, *Reddit Aims for IPO in Second Half as Market's Gears Quietly Turn*, INFORMATION (Feb. 14, 2023), https://perma.cc/XT6C-CS35.

[23] Mike Isaac, *Reddit, Acting Against Hate Speech, Bans "The_Donald" Subreddit*, N.Y. TIMES (June 29, 2020).

the increasingly influential[24] model of politics that seeks a middle ground between, on the one hand, unrealistic hopes for political consensus and, on the other hand, the zero-sum destructiveness of antagonism:

> The bet made by agonism is that by creating a system in which a pluralism of hegemonies is permitted, it is possible to move from an understanding of the other as an enemy, to the other as a political adversary. For this to happen, different ideologies must be allowed to materialize via different channels and platforms. An important prerequisite is that the goal of political consensus must be abandoned and replaced with conflictual consensus. . . . Translated to the Fediverse, it is clear that it already contains a relatively diverse political landscape and that transitions from political consensus to conflictual consensus can be witnessed in the way communities relate to one another. At the base of these conflictual exchanges are various points of view on the collective design and use of the software stack and the underlying protocols that would be needed to further enable a sort of online agonistic pluralism.[25]

The Fediverse is a truly novel evolution in online speech. The question is: It works in theory, but does it work in practice?

II. Content Moderation on the Fediverse

A. The Mastodon Case Study

Although the organization that runs the Mastodon project recommends certain content-moderation policies,[26] each Mastodon instance is able to choose whether and how much to moderate content. The large, general-interest instances tend to have fairly generic policies. For example, Mastodon.social bans "racism, sexism,

---

[24] For a recent attempt to bring agonism into the mainstream of legal scholarship, see Daniel E. Walters, *The Administrative Agon: Democratic Theory for a Conflictual Regulatory State*, 132 Yale L.J. 1 (2022).

[25] Aymeric Mansoux & Roel Roscam Abbing, *Seven Theses on the Fediverse and the Becoming of FLOSS*, *in* The Eternal Network: The Ends and Becomings of Network Culture 124, 131 (Kristoffer Gansing & Inga Luchs eds., 2020). For an influential general account of agonism, see Chantal Mouffe, Agonistics: Thinking the World Politically (2013).

[26] Specifically, the Mastodon project has promulgated a "Mastodon Server Covenant," whereby instances that commit to "[a]ctive moderation against racism, sexism, homophobia and transphobia" such that users will have "confidence that they are joining a safe space, free from white supremacy, anti-semitism and transphobia of other platforms" are eligible to be listed on the project's homepage as recommended instances. *See* Eugen Rochko, *Introducing the Mastodon Server Covenant*, Mastodon (May 16, 2019), https://perma.cc/GP8H-MXXK. But the covenant is not binding on any Mastodon instance, and non-complying instances remain full-fledged member of the overall Mastodon network, subject only to the moderation decision of other instances.

homophobia, transphobia, xenophobia, or casteism" as well as "harassment, dogpiling or doxxing of other users."[27] By contrast, other instances do not specify prohibited categories of content;[28] this, of course, does not prevent the instance administrators from moderating content on an ad-hoc basis, but it does signal a lighter touch. Content moderation can also be based on geography and subject matter; for example, Mastodon.social, which is hosted in Germany, explicitly bans content that is illegal in Germany,[29] and Switter, a "sex work friendly social space" that ran from 2018 to 2022, permitted sex-work advertisements that mainstream instances generally prohibited.[30] Mastodon instances can also impose various levels of moderation on other instances, which can be: (1) fully accessible (the default); (2) filtered but still accessible; (3) restricted such that users can only view content posted on the restricted instances if they follow users on those instances; and (4) fully blocked.

Mastodon instances thus operate according to the principle of content-moderation subsidiarity: Content-moderation standards are set by, and differ across, individual instances. Any given Mastodon instance may have rules that are far more restrictive than those of the major social-media platforms. But the network as a whole is substantially more protective of speech than are any of the major social-media platforms, since no user or content can be permanently banned from the network and anyone is free to start an instance that communicates both with the major Mastodon instances and with the peripheral, shunned instances.

The biggest content-moderation challenge for Mastodon has been Gab, a Twitter-like social network that is popular on the far right. Gab launched in 2016, and, in 2019, switched its software infrastructure to run on a version of Mastodon, in large part to get around Apple and Google banning Gab's smartphone app from their app stores. By switching its infrastructure to Mastodon and operating as merely one of Mastodon's many instances, Gab hoped to hitch a ride back to users' smartphones.[31]

---

[27] *Welcome*, MASTODON, https://perma.cc/326M-JW5A; *see also mas.to!*, MASTODON, https://perma.cc/TBH6-BKWA.

[28] *See, e.g.*, *Mastodon.cloud*, MASTODON, https://perma.cc/7YQQ-ZX87.

[29] *Welcome*, MASTODON, *supra* note 27.

[30] SWITTER, https://perma.cc/B8FA-X7JY.

[31] Adi Robertson, *How the Biggest Decentralized Social Network Is Dealing with Its Nazi Problem*, VERGE (July 12, 2019), https://perma.cc/QA6F-J54U. Gab is not the only right-wing social-

Gab is a useful case study in how decentralized social media can self-police. On the one hand, there was no way for Mastodon to expel Gab from the Fediverse. As Mastodon's founder Eugen Rochko explained, "You have to understand it's not actually possible to do anything platform-wide because it's decentralized. . . . I don't have the control."[32] On the other hand, individual Mastodon instances could—and the most popular ones did—refuse to interact with the Gab instance, effectively cutting it off from most of the network in a spontaneous, bottom-up process of instance-by-instance decisionmaking. Ultimately, Gab was left almost entirely isolated, with more than 99% of its users interacting only with other Gab users. Gab responded by "defederating": voluntarily cutting itself off from the remaining instances that were still willing to communicate with it.[33]

B. Benefits and Drawbacks of Federated Moderation

As the Gab story demonstrates, the biggest benefit of a decentralized moderation model is its embrace of content-moderation subsidiarity: Each community can choose its own content-moderation standards according to its own needs and values, while at the same time recognizing and respecting other communities' content-moderation choices. This is in stark contrast to the problem faced by large, centralized platforms, which by their nature must choose a single moderation standard that different groups of users will inevitably find either under- or overinclusive.

The difference in business models also lowers the need for content moderation generally. The business models of the major platforms—selling advertisements—requires them to maximize "user engagement," and the discovery algorithms designed to promote this goal tend to emphasize conflict across users. By contrast, Fediverse applications can, and often are, engineered with "antivirality" in mind.[34] For example, Mastodon's lack of Twitter's "quote tweet" feature was an intentional design choice on Eugene Rochko's part, who judged that such a feature "inevitably adds toxicity to people's behaviours" and encourages "performative" behavior and

---

media network to use Mastodon as its base. Truth Social, Donald Trump's social-media site, is also built off of Mastodon. Michael Kan, *Trump's Social Media Site Quietly Admits It's Based on Mastodon*, PCMAG (Dec. 1, 2021), https://perma.cc/3CJE-S2AA.

[32] Robertson, *supra* note 31.

[33] Rob Colbert (@shadowknight412), GAB (May 27, 2020), https://perma.cc/G82J-73WX.

[34] Clive Thompson, *Twitter Alternative: How Mastodon Is Designed to Be "Antiviral"*, MEDIUM (Nov. 9, 2022), https://perma.cc/49N4-YWGZ.

"rediculing."[35] The same considerations underpin Mastodon's lack of full-text search and eschewal of algorithmic amplification in favor of reverse-chronological feeds.[36] In addition, Fediverse instances, which are generally run by volunteers and without a profit imperative, can afford to focus on smaller communities in which like-minded users do not suffer the problem of "context collapse" that frequently leads to conflicts on the major social-media platforms.[37]

Of course, if the Fediverse proves popular, for-profit entities may enter the space, thus introducing the problematic incentives of the major platforms. But even if this were to occur, the ability of users to switch Fediverse applications and instances will limit the extent to which the Fediverse's architecture will reflect the values of the extractive attention economy.

The main objection to the Fediverse is that what some see as its key feature—its decentralized model—is for others its main bug. Because there is no centralized Fediverse authority, there is no way to fully exclude even the most harmful content from the network. And, as noted above, Fediverse administrators will generally have fewer resources as compared to giant social-media platforms.[38] By contrast, if Facebook or Twitter want to fully ban a user or some piece of content, they can in principle do so (although in practice it can be a challenge given the size of their networks and users' ability to evade content moderation).

In considering the limits of decentralized content moderation, it is helpful to distinguish between two categories of objectionable conduct. The first category consists of content that is broadly recognized as having no legitimate expressive value. Examples of such content are child-exploitation material, communication that facilitates criminal conduct, and spam. The challenges of moderating these types of content are technological and organizational, and the main question is whether decentralized social media can handle the moderation challenges at scale. Ultimately, it's an empirical question and we'll have to wait until the Fediverse grows to find out the answer. But there are reasons for optimism.

First, the Fediverse itself may be up to the task. Automated scanning, while

---

[35] Eugen Rochko (@Gargron), MASTODON (Mar. 10, 2018), https://perma.cc/VXE7-XVLC.

[36] Thompson, *supra* note 34.

[37] *See, e.g.*, Jenny L. Davis & Nathan Jurgenson, *Context Collapse: Theorizing Context Collusions and Collisions*, 17 INFO., COMMC'N & SOC'Y 476 (2014).

[38] *See supra* notes 9–11 and accompanying text.

hardly foolproof, could lower moderation costs. For example, many of the major platforms use Microsoft's PhotoDNA system to scan for child pornography,[39] and the same software could be used by Fediverse instances for content that they host. And if effective moderation turns out to require more infrastructure, that could lead to a greater consolidation of instances. This is what happened with email, which—in part due to the investments necessary to counter spam—has become increasingly dominated by Google and Microsoft.[40]

If similar scale is necessary to fight spam and bot accounts on the Fediverse, this could serve as a centripetal force to counter the Fediverse's decentralized architecture and lead to a Fediverse that is more centralized than it is today (albeit still far more decentralized than architecturally closed platforms). Partial centralization would reintroduce some of the content-moderation dilemmas that decentralization is meant to avoid,[41] and there is a tradeoff between a vibrant and diverse communication system and the degree of centralized control that would be necessary to ensure 100% filtering of content. The question, to which the answer is as yet unknown, is how stark that tradeoff is.

A second reason to think that federalized systems can have sufficient content moderation is that governments could step in to deal with instances that can't, or choose not to, deal with the worst content. Although the Fediverse may live in the cloud, its servers, moderators, and users are physically located in nations whose governments are more than capable of enforcing local law.[42] A Mastodon instance that hosted child pornography would not only be blocked by all mainstream Mastodon instances, but would also be quickly taken offline—and have its members prosecuted—by the relevant jurisdictions. Even the threat of state action can have large effects. For example, Switter, which by the end of its life was the third-largest Mastodon instance, shut down because its organizers concluded that Switter's continued existence was increasingly untenable as major jurisdictions like the United

---

[39] *See* Hany Farid, *Reining in Online Abuses*, 19 TECH. & INNOVATION 596 (2018).

[40] *See* Enze Liu et al., *Who's Got Your Mail?: Characterizing Mail Service Provider Usage*, *in* PROCEEDINGS OF THE 2021 ACM INTERNET MEASUREMENT CONFERENCE 113 (2021).

[41] For example, the outsize importance of a few email providers has led to complaints of censorship. *See, e.g.*, *Republican National Committee Sues Google over Email Spam Filters*, REUTERS (Oct. 24, 2022), https://perma.cc/49EU-JFEQ.

[42] *See generally* JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET?: ILLUSIONS OF A BORDERLESS WORLD (2006).

States, Australia, and the United Kingdom advanced online-safety and anti-trafficking legislation.[43]

When it comes to the second category of content moderation—content that is objectionable to one group but that others view as legitimate, even core, speech—the Fediverse will host content that current platforms prohibit. But whether this is a weakness or a strength depends on one's substantive views about the content at issue. What looks to one group like responsible moderation can appear to others as unjustified censorship. And when platforms inevitably make high-profile moderation mistakes—moderation, after all, is not an exact science—they undermine their credibility even further, especially where determinations of "misinformation" or "disinformation" are perceived as tendentious attempts to suppress conflict over politics, health, or other important social and culture issues.[44]

The benefit of decentralized moderation is that it can satisfy both those that want to speak and those that don't want to listen. By empowering users, through their choice of instance, to avoid content they find objectionable, the Fediverse operationalizes the principle that freedom of speech is not the same as freedom of reach. In a world where there simply isn't consensus on what content is and is not legitimate, letting people say what they want while giving others the means to protect themselves from that speech may be the best we can do.

A different concern with decentralized moderation is that it will lead to "filter bubbles" and "echo chambers" in which members will choose to only interact with like-minded users.[45] For Mansoux and Abbing, this state of affairs would produce a watered-down, second-best agonism:

---

[43] *See* SWITTER, *supra* note 30.

[44] A high-profile example is Twitter and Facebook's decision on the cusp of the 2020 election to block news reports of Hunter Biden's stolen laptop. While Twitter and Facebook, both of whom played an important role in amplifying Russian election interference in 2016, were understandably concerned that the laptop story was foreign disinformation, later revelations suggesting that the laptop was in fact authentic have further undermined many conservatives' faith in the platforms, and even the platforms themselves have conceded the mistake. *See* Cristiano Lima, *Hunter Biden Laptop Findings Renew Scrutiny of Twitter, Facebook Crackdowns*, WASH. POST (Mar. 31, 2022); Jessica Bursztynsky, *Twitter CEO Jack Dorsey Says Blocking New York Post Story was "Wrong"*, CNBC (Oct. 16, 2020), https://perma.cc/7CMJ-5VGA; David Molloy, *Zuckerberg Tells Rogan FBI Warning Prompted Biden Laptop Story Censorship*, BBC (Aug. 26, 2022), https://perma.cc/XG9Q-5PWQ.

[45] *See generally* CASS R. SUNSTEIN, #REPUBLIC: DIVIDED DEMOCRACY IN THE AGE OF SOCIAL MEDIA (2018).

> Rather than reaching a state of agonistic pluralism, it could be that the Fediverse will create at best a form of bastard agonism through pillarization. That is to say, we could witness a situation in which instances would form large agonistic-without-agonism aggregations only among both ideologically and technically compatible communities and software, with only a minority of them able and willing to bridge with radically opposed systems.[46]

This concern, though understandable, can be addressed several ways. First, filter bubbles are not a Fediverse-only phenomena; closed platforms can design their systems so as to keep dissimilar users from interacting with each other.

Second, it is important to not overstate the effect of filter bubbles; even the most partisan users frequently consume and even seek out information that challenges their beliefs.[47] While Fediverse applications like Mastodon may make it easier for users to communicate only with like-minded peers, users can still go outside their instances to access whatever information they want.

And third, even if filter bubbles exist, it is unclear whether they are a net negative, at least from the perspective of polarization and misinformation. The "backfire effect" (also known as belief perseverance) is a well-established psychological phenomenon whereby individuals who are exposed to evidence that challenge their views end up believing in those views more rather than less.[48] In this view, a more narrowly drawn epistemic environment, while hardly a model of ideal democratic public reason, may actually be better than a social-media free-for-all.

Put another way, the smaller communities of the Fediverse may be a useful corrective to the "megascale" of contemporary social media, which pushes us to "say so much, and to so many, so often."[49]

III. ENCOURAGING THE FEDIVERSE

The Fediverse is still a very small part of the broader social-media ecosystem. Mastodon's several million users pale in comparison with Facebook's billion or Twitter's hundreds of millions of users. Whether the Fediverse ever grows large

---

[46] Mansoux & Abbing, *supra* note 25, at 132.

[47] *See* Peter M. Dahlgren, *A Critical Review of Filter Bubbles and a Comparison with Selective Exposure*, 42 NORDICOM REV. 15 (2021).

[48] *See* Brendan Nyhan & Jason Reifler, *When Corrections Fail: The Persistence of Political Misperceptions*, 32 POL. BEHAV. 303 (2010).

[49] Ian Bogost, *People Aren't Meant to Talk This Much*, ATLANTIC (Oct. 22, 2021), https://perma.cc/U3NT-7MGF.

enough to challenge the current dominance of closed platforms is very much an open question, one that will ultimately depend on whether it provides a product that ordinary users find superior to what is currently available on the dominant platforms.

Such an outcome is hardly preordained. It would require millions of people to overcome the steeper learning curves of Fediverse applications, commit to platforms that are often intentionally less viral than the engagement-at-all-costs alternatives, and navigate the culture shock of integrating into an existing community.[50] After experiencing a mass influx of Twitter users that defected after Elon Musk purchased the platform, Mastodon has seen its active users drop from its late-2022 high of 2.5 million, suggesting that, for many users, Mastodon does not work as a Twitter replacement.[51]

But Mastodon has demonstrated that, for millions of people, decentralized social media is a viable option. And even if Mastodon's market share remains modest, other decentralized applications, whether operating on ActivityPub or other protocols (as with the ATProtocol-powered Bluesky) will continue to grow, especially if they combine Mastodon's emphasis on decentralization with Silicon Valley's engagement-at-all-costs priorities. In the end, the current dominance of the incumbent platforms may prove illusory. They are, after all, themselves subject to shakeups, as is demonstrated by the meteoric rise of apps like TikTok.

Although decentralized social media will have to stand on its own merits, public-policy interventions could nevertheless encourage its growth. Here I briefly consider four such interventions, ranging from most- to least-direct government involvement.

First, governments could support the Fediverse by participating in it as users or, better yet, as instances. This would both directly contribute to the Fediverse's growth but, more importantly, would help legitimate it as the preferred social-media architecture for democratic societies. For example, shortly after Musk announced plans to purchase Twitter, the European Commission, the executive

---

[50] *See* Alan Rozenshtein, *Mastodon's Content-Moderation Growing Pains*, VOLOKH CONSPIRACY (Nov. 21, 2022), https://perma.cc/5MPT-3WYK.

[51] Amanda Hoover, *The Mastodon Bump Is Now a Slump*, WIRED (Feb. 7, 2023), https://perma.cc/TJ5W-YRLZ.

branch of the European Union, launched EU Voice, a Mastodon instance that "provides EU institutions, bodies and agencies with privacy-friendly microblogging accounts that they typically use for the purposes of press and public relations activities."[52] Other governments and international organizations could follow suit.

Second, governments could mandate that large social-media platforms interoperate with the Fediverse. For example, under such a regime, Facebook would be allowed to choose what users or content appear on its servers, but it would have to allow other Fediverse instances to communicate with it. This would allow users to access content that Facebook removes and also still be able to interact with the broader Facebook community.[53] Such regulation would have to specify to what extent Facebook could block other instances entirely, since otherwise Facebook could effectively defederate. But even a limited interoperability mandate would enable a balance between what are the currently envisioned options: totally unfettered control by closed platforms or common-carrier-type regulations that make any sort of moderation impossible.[54]

Such regulation is already being pursued in Europe, where the Digital Services Act would require large platforms to interoperate, a requirement that could easily be modified to include the Fediverse.[55] In the United States, interoperability legislation, which has already been introduced in Congress,[56] would be a welcome alternative to recent overbroad state laws from Texas, Florida, and other Republican-governed states that purport to limit the ability of major social-media platforms to

---

[52] EU VOICE, https://perma.cc/2NTM-9N6E.

[53] Interestingly, Meta is reportedly working on a decentralized text-based social media platform that would interoperate with Mastodon. Deepsekhar Choudhury & Vikas Sn, *Exclusive: Meta Mulls a Twitter Competitor Codenamed "P92" That Will Be Interoperable with Mastodon*, MONEYCONTROL (Mar. 10, 2023), https://perma.cc/6E8L-BFC6.

[54] To be sure, interoperability mandates are not without their own risks, especially to user privacy. *See, e.g.*, Thomas E. Kadri, *Digital Gatekeepers*, 99 TEX. L. REV. 951, 999 (2021); Jane Bambauer, *Reinventing Cambridge Analytica One Good Intention at a Time*, LAWFARE (June 8, 2022), https://perma.cc/7V7W-GML6.

[55] At the same time, other requirements of the Digital Services Act, especially around mandatory content moderation, might hinder the Fediverse's development. *See* Konstantinos Komaitis & Louis-Victor de Franssu, *Can Mastodon Survive Europe's Digital Services Act?*, TECH POL'Y PRESS (Nov. 16, 2022), https://perma.cc/W8RC-2XVL.

[56] *See, e.g.*, Press Release, Sen. Mark R. Warner, *Lawmakers Reintroduce Bipartisan Legislation to Encourage Competition in Social Media* (May 25, 2022), https://perma.cc/SC2Z-3XQL.

moderate content. These laws, in addition to being poorly thought out and overtly political, may also violate the First Amendment, at least in their more extreme versions.[57]

Third, antitrust regulators like the Department of Justice and the Federal Trade Commission could use an incumbent platform's willingness to interoperate as a consideration in antitrust cases.[58] Interoperability could then be an alternative to calls to "break up" social-media giants, a tactic that is both controversial and legally risky.[59]

Finally, policymakers should consider how the background legal regime can be tweaked to improve the incentives for the Fediverse. In the United States, the most important factor is Section 230 of the Communications Decency Act of 1996, which shields platforms from liability as publishers of content created by users.[60] Although Section 230 has come under increasing controversy, especially as it applies to giant platforms, it's hard to imagine how the Fediverse could function without it. The open nature of the Fediverse—with users being able to travel between and communicate across instances—limits the scope of monetization, since users can choose instances that limit advertisements and algorithmic ranking. But this also means that Fediverse instances will lack the resources necessary to perform the sort of aggressive content moderation that would be necessary were they to be held liable for their users' content. The rationale for Section 230 immunity when it was enacted in the mid-90s—to help support a nascent Internet—no longer applies to the technology giants. But it does apply to the current generation of Internet innovators: the federated social-media platforms.

---

[57] *See, e.g.*, Alan Z. Rozenshtein, *First Amendment Absolutism and the Florida Social Media Law*, LAWFARE (June 1, 2022), https://perma.cc/WXT9-4HAL; *see generally* Alan Z. Rozenshtein, *Silicon Valley's Speech: Technology Giants and the Deregulatory First Amendment*, 1 J. FREE SPEECH L. 337 (2021).

[58] *See generally* Chinmayi Sharma, *Concentrated Digital Markets, Restrictive APIs, and the Fight for Internet Interoperability*, 50 U. MEM. L. REV. 441, (2019).

[59] *See* Herbert Hovenkamp, *Antitrust and Platform Monopoly*, 130 YALE L.J. 1952 (2021).

[60] 47 U.S.C. § 230(c)(1).