Friedl, Paul; Morgan, Julian

Article

Decentralised content moderation

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Friedl, Paul; Morgan, Julian (2024) : Decentralised content moderation, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 13, Iss. 2, pp. 1-11, https://doi.org/10.14763/2024.2.1754

This Version is available at:
https://hdl.handle.net/10419/296500

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 13 **Issue** 2

GLOSSARY ENTRY

OPEN ACCESS

PEER REVIEWED

Decentralised content moderation

Paul Friedl *Karlsruhe Institute of Technology*
Julian Morgan *Humboldt Universität Berlin*

DOI: https://doi.org/10.14763/2024.2.1754

Published: 4 April 2024
Received: 10 July 2023 **Accepted:** 25 October 2023

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Friedl, P. & Morgan, J. (2024). Decentralised content moderation. *Internet Policy Review*, *13*(2). https://doi.org/10.14763/2024.2.1754

Keywords: Content moderation, Platform governance, Protocol-based infrastructure, Fediverse

Abstract: Decentralised content moderation describes and potentially advocates for moderation infrastructures in which both the authority and the responsibility to moderate are distributed over a plurality of actors or institutions.

This article belongs to the **Glossary of decentralised technosocial systems**, a special section of *Internet Policy Review*.

Introduction and definition

The last years of academic research have considerably broadened the conventional understanding of online content moderation. It is now largely agreed upon that content moderation takes place across different layers of the technological stack; that there is no hard boundary between content moderation understood as content removal and "softer" forms of moderation such as promotion, demotion, or system design and affordances more broadly; and that content moderation often involves a complex interplay of different actors and activities (Grimmelmann, 2015; Douek, 2022). Despite these newly identified complexities, at its core, content moderation can still be defined as "the screening, evaluation, categorization, approval or removal [...] of online content according to relevant communications and publishing policies" (Flew, Martin, and Suzor, 2019, p. 40).

Within this framework, *decentralised content moderation* describes (and potentially advocates for) moderation infrastructures in which both the authority and responsibility to moderate are distributed over a plurality of actors or institutions. Importantly, rather than denoting a binary opposite to *centralised moderation*, decentralisation only ever indicates a *relative* dispersion of the decision-making power over content propagation.

History and development

Considering that online communication initially occurred mainly through decentralising protocols (e.g. SMTP or TTNP) rather than corporately owned websites, it is unsurprising that in the internet's early days content moderation too was largely decentralised. *Usenet*, a distributed system of threaded messaging boards (so-called *newsgroups*) propagated by a network of independent servers established in 1980, arguably constitutes the most significant example of such early decentralised content moderation practices (see e.g. Philipps, 1996). Whereas *Usenet* itself contained no universal content policy or a central body that could have enforced it, creators of *newsgroups* when setting up a new *newsgroup* could decide to have the group "moderated". In this case, every new user post required prior moderator approval before it would be propagated to other users. This system of prior approval, however, established a bottleneck and critical point of failure that most

creators opted to forgo. Initially, *Usenet's* lack of effective moderation tools remained largely inconsequential: most of its early users belonged to a homogeneous group of academic professionals that had strong incentives to comply with a common netiquette so as not to endanger their in-group reputation. However, once more varied demographics gained private internet access, such discursive conventions were hard to maintain. Spam emerged as another immense problem, seriously endangering the service's usability (Lucas & Miller, 2023). *Usenet* servers' ability to ban individual users, only carry newsgroups with certain credentials, or to not propagate the content from other servers deemed to carry excessive amounts of spam, harassment, or other "bad" content, provided only rather crude tools to mitigate these problems (Lucas & Miller, 2023). Not least for this reason, *Usenet*, while still existing, has gradually lost its relevance.

Web 2.0 then saw the emergence and rise of much more centralised fora of discussion, i.e. what are still today seen as the dominant social networks such as *Twitter* (now called *X*), *Facebook*, *Instagram* or *TikTok*. Most of these platforms have implemented uniform, platform-wide content policies, which are enforced by largely centralised moderation apparatuses. However, even in this era of "walled gardens'' some services have retained more community-driven, decentralised forms of content governance. *Reddit*, where users discuss topics in interest-related message boards (so-called *subreddits*), arguably constitutes the most significant example of this form of decentralised content moderation. Although *Reddit* has established an (arguably rather vague) site-wide content policy, which its employed so-called *administrators* can enforce e.g. by removing content, suspending accounts or deleting entire *subreddits*, the absolute lion's share of moderation is performed by *subreddits'* own moderators (Newton, 2023). Conventionally, moderators are either the users who created a *subreddit* or other users appointed by these creators. Moderators not only need to uphold *Reddit's* site-wide rules but are also encouraged to set down and enforce *subreddit*-specific rules (Reddit, 2023). Attempting to render discussions as on-topic and engaging as possible, these rules frequently go far beyond banning "harmful" content and often prescribe in detail what posts and comments are allowed. As it is often individuals who participate in the discussion themselves that moderate *subreddits*, this form of user-governed moderation is sometimes also called (community) self-moderation (Seering, 2020). Another, albeit more limited, example of community moderation is provided by *X's*) "community notes" feature. "Community notes", intended as a crowd-sourced fact-checking tool, lets users add additional information to the posts of other users (such as a note evidencing the untruthfulness of the original post) if enough users with "diverse perspectives" have rated the additional information as helpful (Masnick,

2023; Wirtschafter & Majumder, 2023). Other largely centralised platforms, too, have recently started to partially rely on moderation crowd-sourcing, such as by introducing "trusted flaggers" (YouTube, Facebook, TikTok) or group self-moderation (Discord), arguably sometimes with an eye to cutting "trust and safety" expenses and outsourcing difficult moderation decisions.

Finally, in an attempt to provide netizens with alternatives to the dominance of large social media sites, there has recently been renewed interest in and intense development of genuinely decentralised online communication infrastructures, spearheaded by the *ActivityPub* protocol and its flagship social media service *Mastodon*. Much like *Usenet*, *Mastodon* does not constitute one single website but rather consists of an ensemble of networked (*federated*) independent servers (so-called *instances*). Each *instance* is free to determine its own local rules. It can also choose to block other *instances*, rendering content posted to such *instances* invisible to its own users. However, there is no superior central institution able to ban or delete individual *instances* or impose *instance*-transcending content policies. As with the other mentioned examples of distributed content moderation, users generally cannot directly influence an *instance's* content policies. Moreover, in July 2023, *Meta* launched *Threads*, a *Twitter*-inspired microblogging service, which it promised would soon be compatible with *Activity Pub* (Meta, 2023). Were this to happen, *Threads* would constitute a true hybrid: an instance on a decentralised network that itself exhibits a greatly centralised form of (content) governance. *Bluesky*, another new microblogging service developed on top of a new decentralised networking protocol called the *AT protocol*, constitutes a very similar example. Next to the decentralisation of moderation enabled by the protocol's federation-based architecture, *Bluesky* has also promised to give users more power to individualise moderation through features such as customisable recommender engines and other moderation "middleware" (see also Masnick, 2019; Keller, 2021).

Typology

As was already mentioned above, decentralisation, as a sociotechnical process, cannot be understood in binary terms. Rather, it must necessarily be expressed in terms of degrees to which a system or infrastructure is distributed over different 'nodes' of decision-making power (Bodó et al., 2021; Sai et al., 2021). Furthermore, as the preceding historical overview has shown, decentralised content moderation can consist of a variety of practices and takes on many different forms. Despite this heterogeneity, it is possible to typify different forms of decentralised content moderation. In particular, it is possible to distinguish three different factors contribut-

ing to the overall decentralisation of a given content moderation system: the allocation of jurisdiction/authority between different nodes, the relative autonomy of any node of governance (with respect to other, especially superior nodes of governance) as well as the ease of exit or migration. All three of these factors have a decisive impact on the degree of (de)centralisation of a given content moderation system.

The first axis of classification relates to the formal aspect of *jurisdiction*[1], i.e. which and how many governance nodes have decisional power or authority over a specific piece of content. Different nodes of content governance may either have exclusive or overlapping jurisdiction over some piece of content. Overlapping jurisdiction has the effect of decentralising the decisional power over content to the different nodes. Large social media platforms, for instance, generally have exclusive universal jurisdiction over most user-generated content, meaning that the company's own team of moderators is the only governance node that can delete content for all users. At the same time, however, company moderation teams might share jurisdiction over content posted in groups (e.g. on *Facebook*) with the administrators of these groups. A similar case of overlapping jurisdictions may be observed on *Reddit*, where subreddit administrators' power to moderate a certain subreddit does not preclude the concurrent ability of Reddit's own 'Trust and Safety' team to take moderation actions regarding such contents.

Second, iterations of decentralised content moderation can be typified based on the *autonomy* of content moderation nodes from the corresponding authority of other, especially hierarchically superior nodes of governance. For instance, from one perspective, one might say that the moderation practices of large social media platforms are heavily decentralised as decisions are spread out over enormous groups of human moderators, often also involving independent contractors (Klonick, 2018, p. 1639). Ultimately however, whereas these systems clearly involve a degree of multi-polar or multi-level governance, the specificity and rigidity of internal moderation guidelines will rarely leave moderators with substantial decisional autonomy, resulting in a moderation system that is *effectively* centralised (Gray, 2022). Another important factor in gauging moderation autonomy is whether decision-making powers are technologically entrenched or not. For instance, while *Reddit* grants *subreddit* moderators substantial moderation powers, these "privileges" could easily be revoked, e.g. in case *Reddit* were to decide to change the site's structure or rely more heavily on centralised content moderation

1. We use the word "jurisdiction" in a non-legalistic manner to indicate whether a specific node has adjudicatory power and authority on a given piece of online content.

(Peters, 2023). Conversely, content governance structures that are hardcoded in technological protocols, such as on *Usenet*- or *ActivityPub*-based platforms, grant decentralised moderation nodes much greater security and autonomy, guaranteeing that their moderation practices will not be undermined or meddled with from above (Pierce, 2021). In other words, they consolidate autonomy at the infrastructural level (Ermoshina, 2022, p. 6).

Third, the last factor that determines a platform's effective level of moderation decentralisation is the ability of and relative ease with which participants may exit or migrate to new forums (Hirschman, 1972). The easier and less costly it is for forum administrators or normal users to create or move to a new forum, the more likely it is for a platform to actually decentralise and offer users multiple fora of discussion. Neither *Twitter (now "X")*, *Instagram* nor *TikTok*, for example, offer users any possibilities to choose between different fora/moderation systems, thus precluding one very fundamental level of potential decentralisation. *Reddit* or *Mastodon*, on the other hand, do allow users to create or migrate to a new *subreddit* or *instance*. What might be different between these last two examples then are the relative costs of migration. Factors that determine this cost include the technical ease with which users might switch to a different forum or whether they risk losing the fruits of their past platform engagement, such as existing followers or other credentials.

Normative trade-offs

Both centralised and decentralised forms of content moderation possess distinct strengths and weaknesses. The perhaps biggest advantage of decentralised moderation lies in its ability to allocate moderation responsibilities to those most intimately aware of the intricacies of specific online communities. Moderators are thus frequently deeply invested in the objectives, health and flourishing of these communities and garner specific experience and know-how on how to achieve these objectives. Grounded in the principles of conventional federalism as a means to restrain and contain power while fostering systemic efficiency, pluralism, expertise and individual liberty, decentralised content moderation is a potential response to many of the issues linked to the current prevalent models in the advertisement-driven attention economy (Kadri, 2022, p. 198; Halberstam, 2012). Distributed community self-moderation can also lead to more participatory, democratic forms of online governance (Bietti, 2021, p. 62). Finally, decentralised content moderation can ensure discursive diversity across the entire network and can also mitigate the emergence of excessive moderation power or 'single points of failure' (Ermoshina, 2022, p. 2).

Decentralising content moderation, however, also poses risks. For one thing, decentralised moderation may create so-called echo chambers and exacerbate discursive fragmentation and societal polarisation (Dubois & Blank, 2018; Cineli et. al, 2021). Of course, whether such online echo chambers indeed are undesirable and whether they can be effectively prevented are still largely open questions (Rozenshtein, 2023). Other disadvantages of decentralised content moderation, however, are more tangible. First among these is the issue that moderation is resource- and time-intensive (Gillespie, 2018). Decentralised moderation efforts may not only lack the ability to efficiently scale - as centralised systems do rather easily, but may also suffer from the unavailability of funding streams. Indeed, the ideals of sustainable community self-moderation may be at odds with the profit-driven logic necessary to access conventional funding streams (Rozenshtein, 2023).

This lack of funding streams and the risk of chronic overburdening of moderation teams also means that additional user safeguards, such as greater transparency or appeal mechanisms, may sometimes be non-viable. What is more, certain concerted attacks against online fora, such as covert infiltration with bots, may be detectable only by system administrators; decentralised moderation teams will often lack the data and tools to counteract such activities. One solution to this operational hurdle could lie in greater and more accessible use of automated moderation tools, as advances in machine learning technologies are reducing the costs of detecting and removing undesirable content (Jhaver et al., 2019). In any case, due to the speed, scale, and resources required for such moderation, the risk of commercial capture of content moderation in decentralised systems could lead to forms of effective re-centralisation of moderation practices (Ermoshina, 2022, p. 14).

The impact of law on decentralised content moderation

The legal landscape relative to decentralised content moderation is complex and dynamic (see e.g. Mazgal, 2021). It is still largely unclear if and, if so, to what extent decentralised moderation nodes, such as *subreddit* moderators or *Mastodon* instance administrators, fall under existing liability exemptions for content intermediaries, such as section 230 of the US Communications Decency Act (Ahooja, 2023). In the European Union, from January 2024 onwards, these liability questions and other issues of regulated platform governance are governed by the EU's new Digital Services Act (see generally, Eifert et al., 2021). Here too, it is unclear if and, if so, which forms of decentralised hosting and moderation actors will be qualified as

"providers" of an "intermediary service" and thus benefit from the Act's liability exemptions. It is furthermore unclear if and, if so, which forms of decentralised hosting and moderation actors will be qualified as "online platforms" and will thus need to comply with the DSA's demanding array of user-protection obligations (e.g. the obligation to provide an "internal complaint-handling system" through which recipients can obtain review of specific moderation decisions or the obligation to publish yearly transparency reports) (Komaitis & De Franssu, 2022). Arguably incidental aspects, such as which actor has ultimate control over the technical infrastructure or which actor is legally responsible for hosting certain online contents, might end up determining these designations and lead to potentially arbitrary and inconsistent results (e.g. *Mastodon* instances being classified as "providers", while *Reddit* or *Discord* forums being spared such treatment).

Unfortunately, the EU seems to have failed to fully consider or plan for the impacts its new platform law might have for decentralised content moderation. This fits into a broader picture: regulators and policymakers the world over have so far largely ignored the special needs and qualities of decentralised moderation initiatives and have thus also neglected the potentials and opportunities offered by this form of online community governance (Mazgal, 2021; Keller, 2021). Policymakers should pay more attention to decentralised content moderation when crafting online governance regimes. One way of fostering distributed forms of moderation could lie in imposing interoperability mandates to large platforms similar to those enshrined in the EU's Digital Markets Act (Rozenshtein, 2023). In any event, lawmakers should adopt a broad perspective when regulating online moderation instead of considering only the few large platforms, which risks resulting in potentially insurmountable barriers to entry for smaller (decentralised) online communities.

Conclusion

There can be little doubt that online communication, like all communication, is most likely to serve and satisfy its participants, where it happens in a spirit of respect and cooperativity, with all participating parties ideally sharing a common purpose. By decentralising content moderation, online communities are incentivised to reflect, debate and agree on the purposes of their community, which can instil a virtuous *culture of responsibility* on participants and moderators alike. Content moderation and community governance should also be informed by the conventional federalist principles of *subsidiarity and localism*, and especially the impetus that governance should generally take place at the lowest level unless otherwise justified (e.g. by motives of practicality) (Blank, 2010; Kadri, 2022). The show-

cased examples of decentralised moderation, however, also show how important it is that a platform or protocol provides moderators with adequate tools and sociotechnical structures to perform their tasks effectively and efficiently. Moreover, superior nodes of governance will likely not be able to delegate activities related to the prevention and mitigation of structural issues and attacks. Thus placing an emphasis on the need to design systems with a priority on creating healthy discursive spheres. This may mean giving up on prioritising engagement and growth practices that go against user interests and communities' health. It also underlines the importance of meaningful migration possibilities for users that guarantee that community moderation remains responsive to community needs.

References

Ahooja, R. (2023). *Section 230 and the Fediverse: The 'instances' of Mastodon's immunity and liability*. SSRN. https://doi.org/10.2139/ssrn.4421665

Bietti, E. (2023). A genealogy of digital platform regulation. *Georgetown Law Technology Review*, *7*(1), 1–68. https://doi.org/10.2139/ssrn.3859487

Blank, Y. (2010). Federalism, subsidiarity, and the role of local governments in an age of global multilevel governance. *Fordham Urban Law Journal*, *37*(2), 509–558. https://ir.lawnet.fordham.edu/ulj/vol37/iss2/1

Bodó, B., Brekke, J. K., & Hoepman, J.-H. (2021). Decentralisation: A multidisciplinary perspective. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1563

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), Article e2023301118. https://doi.org/10.1073/pnas.2023301118

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729–745. https://doi.org/10.1080/1369118X.2018.1428656

Eifert, M., Metzger, A., Schweitzer, H., & Wagner, G. (2021). Taming the giants: The DMA/DSA package. *Common Market Law Review*, *58*(4), 987–1028. https://doi.org/10.54648/cola2021065

Ermoshina, K., & Musiani, F. (2022, November). *Safer spaces by design? Federated architectures and alternative socio-technical models for content moderation*. Annual Symposium of the Global Internet Governance Academic Network (GigaNet). https://hal.science/hal-03930548/document

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, *10*(1), 33–50. https://doi.org/10.1386/jdmp.10.1.33_1

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://doi.org/10.12987/9780300235029

Gray, C. (2022). *The moderator: Inside Facebook's dirty work in Ireland*. Gill Books.

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law & Technology*, *17*, 42–109. ht tp://hdl.handle.net/20.500.13051/7798

Hirschman, A. O. (1972). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard University Press.

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of Reddit automoderator. *ACM Transactions on Computer-Human Interaction*, *26*(5), 1–35. https://doi.org/10.1145/3338243

Kadri, T. E. (2022). Juridical discourse for platforms. *Harvard Law Review*, *136*(2), 163–204.

Keller, D. (2021). The future of platform power: Making middleware work. *Journal of Democracy*, *32*(3), 168–172. https://doi.org/10.1353/jod.2021.0043

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*(6), 1598–1670.

Komaitis, K., & de Franssu, L.-V. (2022, November 16). Can Mastodon survive Europe's Digital Services Act? *Tech Policy Press*. https://techpolicy.press/can-mastodon-survive-europes-digital-servi ces-act/.

Lucas, R., & Miller, T. (2023). *Federation and moderation: Usenet as the original decentralised social network* [Video]. LibrePlanet. https://media.libreplanet.org/u/libreplanet/m/federation-and-moderat ion-usenet-as-the-original-decentralized-social-network/.

Masnick, M. (2019). Protocols, not platforms: A technological approach to free speech. *Knight First Amendment Institute*, *19*(5). https://perma.cc/MBR2-BDNE

Masnick, M. (2023, October 31). Community notes is a useful tool for some things… But not as a full replacement for trust and safety. *Techdirt*. https://www.techdirt.com/2023/10/31/community-notes-i s-a-

Mazgal, A. (2021). Back to the future? The Digital Services Act and regulating online platforms built on community-led moderation. In A. Baratsits (Ed.), *Building a European digital public space. Strategies for taking back control from big tech platforms*. iRights Media. http://dx.doi.org/10.2139/ssr n.3919102

Newton, C. (2023, June 13). How Reddit set itself up for a fall. *The Verge*. https://www.theverge.com/ 2023/6/13/23759130/reddit-protests-history-community-growth-moderation

Peters, J. (2023, June 30). How Reddit crushed the biggest protest in its history. *The Verge*. https://w ww.theverge.com/23779477/reddit-protest-blackouts-crushed

Philips, D. (1996). Defending the boundaries: Identifying and countering threats in a Usenet newsgroup. *The Information Society*, *12*(1), 39–62. https://doi.org/10.1080/019722496129693

Pierce, D. (2023, April 20). Can ActivityPub save the internet? *The Verge*. https://www.theverge.com/ 2023/4/20/23689570/activitypub-protocol-standard-social-network

Reddit. (n.d.). *Moderator code of conduct*. https://www.redditinc.com/policies/moderator-code-of-con duct

Rozenshtein, A. Z. (2023). Moderating the Fediverse: Content moderation on distributed social media. *Journal of Free Speech Law*, *3*, 217–236. https://doi.org/10.2139/ssrn.4213674

Sai, A. R., Buckley, J., Fitzgerald, B., & Le Gear, A. (2021). Taxonomy of centralization in public

blockchain systems: A systematic literature review. *Information Processing & Management*, *58*(4), Article 102584. https://doi.org/10.1016/j.ipm.2021.102584

Seering, J. (2020). Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–28. https://doi.org/10.1145/3415178

Wirtschafter, V., & Majumder, S. (2023). Future challenges for online, crowdsourced content moderation: Evidence from Twitter's community notes. *Journal of Online Trust and Safety*, *2*(1), 1–11. https://doi.org/10.54501/jots.v2i1.139

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

cnrs centre internet et societe

UOC R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies