TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY-SA

Citation (APA)
Gsenger, R., Kübler, J., & Wagner, B. (2022). Content moderation on social networking sites: Battling disinformation and upholding values. In *Entscheidungsträger im Internet: Private Entscheidungsstrukturen und Plattformregulierung* (pp. 181-200) https://doi.org/10.5771/9783748934981

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Content governance on social networking sites: Battling disinformation and upholding values

Rita Gsenger, Johanne Kübler, Ben Wagner

I. Introduction[1]

Social media platforms such as Facebook, YouTube, and Twitter have millions of users logging in every day, using these platforms for communication, entertainment, and news consumption. These platforms adopt rules that determine how users communicate and thereby limit and shape public discourse.[2]

Platforms need to deal with large amounts of data generated every day. For example, as of October 2021, 4.55 billion social media users were active on an average number of 6.7 platforms used each month per internet user.[3] As a result, platforms were compelled to develop governance models and content moderation systems to deal with harmful and undesirable content, including disinformation. In this study:

- 'Content governance' is defined as a set of processes, procedures, and systems that determine how a given platform plans, publishes, moderates, and curates content.
- 'Content moderation' is the organised practice of a social media platform of pre-screening, removing, or labelling undesirable content to reduce the damage that inappropriate content can cause.

Online platforms rely on content moderation to guarantee their compliance with laws and regulations, community guidelines, and user agreements, as well as norms of appropriateness for a given site and its cultural

1 Parts of this article are based on an earlier research project funded by the European Green Party, the results of which can be accessed here: https://www.greens-efa.eu/files/assets/docs/alternative_content_web.pdf.

2 Suzor/Van Geelen/Myers West, Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda, International Communication Gazette 2018, (385) 386.

3 Kemp, Digital 2021 October Global Statshot Report, 2021, https://datareportal.com/reports/digital-2021-october-global-statshot.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

context.[4] While content moderation helps to keep undesirable content at bay, discussions are held on the influence of platforms on freedom of expression and information, including individuals' digital rights. The Covid-19 pandemic has increased the urgency of these discussions, and public health policies are influenced by disinformation on social media platforms. It has even led public policymakers to refer to a Covid-related ''infodemic'' to describe the degree of disinformation being spread on social media platforms.[5]

As public attention to the role of false information on online platforms increased, the academic exploration of disinformation and related terms such as misinformation, malinformation, and ''fake news'' greatly expanded. Definitions of disinformation abound, however, they usually include a deliberate intent to cause harm by disseminating false information. Disinformation can thus be defined as the purposeful dissemination of erroneous information with the goal to influence public opinion, groups, or individuals for political or economic gain.[6] In contrast to misinformation, whose inaccuracies are unintended, disinformation is deliberately false information spread intentionally. The use of another term in the same universe, “fake news”, while initially explored by academics,[7] has been largely dismissed as a political expression used to criticise news stories or media outlets.[8]

Research indicates that while dis- and misinformation make up a relative small portion of all information shared on online platforms,[9] their

4 Roberts, Behind the screen: content moderation in the shadows of social media, 2019, 33 ff.

5 Cinelli/Quattrociocchi/Galeazzi/Valensise/Brugnoli/Schmidt/Zola/Zollo/Scala, The COVID-19 social media infodemic, Scientific Reports 2020, 10 (16598), […], 1 (1).

6 Wardle/Derakhshan, Information Disorder: Toward an interdisciplinary framework for research and policy making, Council of Europe Report DGI, 2017, […], 1 (15 ff).

7 E.g. Allcott/Gentzkow, Social Media and Fake News in the 2016 Election, Journal of Economic Perspectives 2017, 211 (212).

8 European Commission, Directorate-General for Communication Networks, Content and Technology, A multi-dimensional approach to disinformation: report of the independent high level group on fake news and online disinformation, Publications Office, 2018, 1 (5).

9 Guess/Nagler/Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook, Science Advances 2019, 5(1), 1 (5); Kübler/Sekwenz/Rachinger/König/Gsenger/Pirkova/Wagner/Kettemann/Krennerich/Ferro, The 2021 German Federal Election and Social Media: Studying the prevention of systemic electoral risk based on the EU Digital Services Act, Report, 2021, 1 (27 ff.).

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

use can be strategic in contexts such as elections.[10] A considerable amount of research focuses on the identification of disinformation,[11] and the motivation of individuals to believe and spread disinformation.[12] Various counter mechanisms have been suggested such as content warnings[13] and fact checking[14], detection of disinformation,[15] automated recognition[16], and deletion of undesired content[17]. Many legislative approaches focus

10 Howard/Kollanyi, Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum, 2016, https://ssrn.com/abstract=2798311.

11 Qian/Gong/Sharma/Liu, Neural User Response Generator: Fake News Detection with Collective User Intelligence, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, 3834 (3835).

12 Scott/Kosslyn, Emerging Trends in the Social and Behavioral Sciences/Cook/Ecker/Lewandowsky, 2015, 1 (8 f.); Islam/Laato/Talukder/Sutinen, Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective, Technological Forecasting and Social Change 2020, 159, 1 (5); Buchanan/Benson, Spreading Disinformation on Facebook: Do Trust in Message Source, Risk Propensity, or Personality Affect the Organic Reach of "Fake News"?, Social Media + Society 2019, 5(4), 1 (4 ff.); Buchanan, Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation, PLOS ONE 2020, 15(10), 1 (14 f.).

13 Kaiser/Wei/Lucherini/Lee, Adapting Security Warnings to Counter Online Disinformation, 30th USENIX Security Symposium, 2021, 1163 (1166 ff.).

14 Vlachos/Riedel, Fact Checking: Task definition and dataset construction, Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science 2014, 18 (19 ff.); Clayton et al., Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media, Political Behavior 2020, 1073 (1083 ff.).

15 Tschiatschek/Merchant/Singla/Krause/Gomez Rodriguez, Fake News Detection in Social Networks via Crowd Signals, WWW 2018, 517 (518 f.); Kim/Tabibian/Oh/Schölkopf/Gomez-Rodriguez, Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation, WSDM 2018, 324 (327 ff.).

16 Yankoski/Weninger/Scheirer, An AI early warning system to monitor online disinformation, stop violence, and protect elections, Bulletin of the Atomic Scientists 2020, 76 (2), 85 (85 f.); Della Vedova/Tacchini/Moret/Ballarin/DiPierro/de Alfaro, Automatic Online Fake News Detection Combining Content and Social Signals, 22nd Conference of Open Innovations Association (FRUCT), 2018, 272 (274); Seo/Xiong/Lee, Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation, Proceedings of the 10th ACM Conference on Web Science, 2019, 265 (267); Alaphilippe/Gizikis/Hanot/Nomtcheva, Automated tackling of disinformation: major challenges ahead, European Parliament Directorate-General for Parliamentary Research Services, 2019, 1 (27 ff.).

17 Bastos, Five Challenges in Detection and Mitigation of Disinformation on Social Media, 2021, https://ssrn.com/abstract=3874410.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

on the latter, making deletion online platforms' most common content moderation approach. Large social media platforms such as Facebook and Twitter use fact-checking and labelling content strategies to counter disinformation, and if needed, deletion of content.[18] However, content governance models focusing on deletion struggle to effectively moderate content that is difficult to classify, such as disinformation. In fact, due to short time frames for content removals, the threat of heavy fines for non-compliance and legal uncertainty, platforms frequently over comply and remove online content without transparency or inclusion by a larger public.[19]

Alternatives to content removal exist, and they are applied successfully by community-led platforms. These platforms are governed partially or entirely by their users and are primarily small in numbers. Many large platforms do not use these methods of content moderation, and if they do, only after significant public pressure.[20] Alternative approaches often focus on strengthening the community to decrease the necessity of content moderation.[21] However, moderators on these community-led platforms are usually volunteers and thus struggle with problems of disinformation. For online communities to function effectively and thrive, many disruptions such as rule-breaking need to be avoided, which we argue is possible through design and compelling content governance.

To assess the effectiveness and value of alternative content governance models practices, this study closely investigates three community-led platforms, namely mastodon in Section →IV.1., diaspora* in Section → IV.2., and Slashdot in Section → IV.3. This investigation is complemented with interviews with experts and administrators of said platforms. Based on the outcomes, we develop an analysis of advantages and disadvantages for community-based and user-centric platform administration in Section →V.

---

18 Iosifidis/Nicoli, The battle to end fake news: A qualitative content analysis of Facebook announcements on how it combats disinformation, International Communication Gazette 2020, 81(1), 60 (69 ff.); Bastos, Five Challenges in Detection and Mitigation of Disinformation on Social Media, 2021, https://ssrn.com/abstract=3874410.

19 Ahlert/Marsden/Yung, How 'Liberty' Disappeared from Cyberspace: The Mystery Shopper Tests Internet Content Self-Regulation, PCMLP Research Paper 2014, 1 (26 f.).

20 Goldman, Content Moderation Remedies, Michigan Technology Law Review 2021, 1 (26).

21 Lampe/Resnick, Slash(dot) and burn: distributed moderation in large online conversation space, Proceedings of the 2004 conference on Human factors in computing systems- CHI '04, 2004, 543 (545 ff.).

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

II. Theoretical framework

Just like real-life communities condone some types of behaviour and reject others, so do communities on online platforms. For instance, Wikipedia requires its authors to remain neutral to further the goal of creating a trustworthy encyclopaedia.[22] These norms can be implicit or take a written form, and they may be contested and change over time. However, most of them need to be accepted by the most significant part of the community to be effective.

This consensus helps deal with the inevitable conflicts and disruptions occurring on online platforms, created by both platform insiders and outsiders. Conflicts from inside the platform might stem from newcomers, who are unaware of or disagree with some norms.[23] Disruptions from outsiders, on the other hand, include trolling—provocative, irrelevant, or attention seeking posts aiming to provoke emotional response.[24] Outsiders such as trolls are relatively immune to sanctions because they are not invested in the platform compared to insiders.[25] Persistent violations of behavioural norms and protracted conflicts can cause serious damage to online communities, which can be averted by adopting measures to decrease the frequency of non-normative behaviours or lessen their impact on the community.

The theoretical framework of this article is primarily based on Regulating Behavior in Online Communities by Kiesler et al. (2012)[26], Governing Internet Expression (Wagner 2016)[27], Custodians of the Internet by Gillespie (2018)[28], Behind the Screen by Roberts (2019)[29], and Content

22 Wikipedia, the free encyclopedia, Wikipedia: Neutral point of view, 2022, https://en.wikipedia.org/ wiki/Wikipedia:Neutral point of view.

23 Kraut/Resnick, Building Successful Online Communities: Evidence-Based Social Design/Kiesler/Kraut/Resnick, 2012, 125 (129).

24 Schwartz, The Trolls Among Us, The New York Times, 2008, https://www.nytimes.com/2008/08/03/magazine/03trolls-t.html.

25 Kraut/Resnick, Building Successful Online Communities: Evidence-Based Social Design/Kiesler/Kraut/Resnick, 2012, 125 (127).

26 Kraut/Resnick, Building Successful Online Communities: Evidence-Based Social Design/Kiesler/Kraut/Resnick, 2012, (125) 125.

27 Wagner, Global Free Expression: Governing the Boundaries of Internet Content, 2016.

28 Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media, 2018.

29 Roberts, Behind the screen: content moderation in the shadows of social media, 2019.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

Moderation Remedies by Goldman (2021)[30]. Through choices in designing platforms, it is possible to: (1) limit the extent of damage caused by bad behaviour; (2) control how much bad behaviour that any bad actor can engage in to begin with; and, (3) encourage voluntary compliance with norms through various incentives. These measures can be combined to increase their effectiveness.

In fact, platforms are legally obliged to remove content that is "manifestly unlawful", according to EU law (outlawing child abuse material,[31] racist and xenophobic hate speech,[32] terrorist content,[33] and infringing intellectual property rights[34]) and various provisions by individual EU Member states, for instance, the German Netzwerkdurchsetzungsgesetz or the French Loi Avia.[35] The proposed Digital Services Act should increase platforms' accountability for content by obligating them to regularly conduct a risk assessment and take appropriate mitigation measures.[36] In addition, platforms also moderate content in contravention of their own Terms of Service (ToS), to which a person must agree to abide by when registering an account, and their community standards, which govern the behaviour of all community members during their participation in that community, setting out which content and behaviours are deemed unacceptable, disruptive, or inappropriate.[37]

30 Goldman, Content Moderation Remedies, Michigan Technology Law Review 2021, 1 (1).

31 Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography and replacing Council Framework Decision 2004/68/JHA, Art. 25.

32 Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online.

33 Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA, Art. 21.

34 DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

35 Hoffman/Gasparotti, Liability for illegal content online Weaknesses of the EU legal framework and possible plans of the EU Commission to address them in a "Digital Services Act", 2020, 1 (25).

36 European Commission/Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, Art. 26 and Art. 27.

37 Roberts, Behind the screen: content moderation in the shadows of social media, 2019, 69.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

In order to make the moderation process on a platform more effective, it is paramount that decisions by content moderators are perceived as legitimate; otherwise, they can lead to more disruption.[38] A consistent application of moderation criteria, giving offenders the possibility to argue their case with the moderator as well as appeal procedures,[39] increases the legitimacy of the moderation process and thus the effectiveness of moderation decisions. Acceptance of the moderation process can furthermore be increased by avoiding the removal of inappropriate content, instead redirecting it to other places. Finally, moderation gains legitimacy when moderators are members of the community perceived as impartial and endowed with limited or rotating powers.[40] Research on automated tools in content moderation has found that their tendency to make mistakes threatens to erode users' trust in the moderation process. When employed carefully in conjunction with human content moderation, however, they can be useful. Bots and other automated tools can be employed, for example, to provide explanations for content removal, helping to prevent future post removals.[41]

Besides moderating the bad behaviour of individual platform users, design choices can help to limit possible destructive behaviour. These include throttles or quota mechanisms to prevent repetitive spam-like activity in a short time frame, or striking continuous disrupters with gags or bans. Some communities have developed internal currencies or ladders of access to force members to earn certain privileges before they can engage in potentially harmful activities. Currency and thus privileges can be gained through everyday participation, such as providing genuine information, which is easy for normal platform participants, but difficult for trolls and manipulators.[42]

Finally, insiders, who care about the community, can be induced to comply voluntarily by instituting behavioural norms. Platform users infer

38 Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media, 2018, 8.

39 Jhaver/Birman/Gilbert/Bruckman, Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator, ACM Transactions on Computer-Human 2019, 26(5), 1 (13 f.).

40 Kraut/Resnick, Building Successful Online Communities: Evidence-Based Social Design/Kiesler/Kraut/Resnick, 2012, 125 (132).

41 Jhaver/ Birman/Gilbert/Bruckman, Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator, ACM Transactions on Computer-Human 2019, 26(5), 1 (22).

42 Kraut/Resnick, Building Successful Online Communities: Evidence-Based Social Design/ Kiesler/Kraut/Resnick, 2012, 125 (131 ff.).

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

norms through the observation of other actors and of consequences to their behaviour, seeing instructive rules of conduct, and participating and receiving feedback. Through design choices, platform creators can communicate normative and thus desired behaviours, for example, by conspicuously highlighting instances of desired behaviour and inappropriate behaviour when participants may be about to violate them.[43] As awareness of norms does not guarantee their adoption, including the community in drafting rules increases their legitimacy and hence compliance. Reputation systems, which summarise a participant's past online behaviours into a score, further increase norm compliance.[44] Finally, norm violations may be prevented through authentication of identities or through incentives to retain a long-term identifier in communities relying on pseudonyms.[45]

III. Research design

To investigate alternatives to content moderation that go beyond deletion, this study presents three case studies of community-led platforms that are unusual cases.[46] These platforms highlight ways of strengthening online communities and thereby decreasing the necessity of content moderation.

Furthermore, 16 semi-structured interviews were conducted in January and February 2021 with content moderators, platform administrators, and experts. 'Experts' refers to researchers studying online platforms and their governance from various disciplines such as law, communications, and social sciences. Table 1 shows an overview of participants including the type of expertise and participants' years of experience in that area. Participants were anonymised and are referred to as P1-P16 throughout the study.

43 Leader Maynard/Benesch, Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention, Genocide Studies and Prevention, 2016, 9(3), 70 (85 f.).

44 Goldman, Content Moderation Remedies, Michigan Technology Law Review 2021, 1 (40).

45 Caplan, Content or Context Moderation: Artisanal, Community-Reliant and Industrial Approaches, 2018, 1 (20 f).

46 Jahnukainen, Extreme Cases/Mills/Durepos/Wiebe, 2010, 378 (378).

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

Table 1: Summary of participants' expertise and years of experience.

Type of Expertise	**No.**
Research	6
Industry/NGO	5
Moderator/Administrator	4
Years of Experience	**No.**
<5	1
5-10	7
>10	4
N/A	3

The interviews were held in a non-directed manner to guide the conversation and allow the participants to tell their stories uninterrupted.[47] Due to safety concerns during the Covid-19 pandemic, the interviews were entirely held online.[48] Even though videoconference is the most suitable alternative to face-to-face interviews,[49] participants might have trouble with the technology or a poor internet connection might make the conversation challenging.[50] Therefore, written responses to the interview questions were

47 McCracken, The Long Interview, 1988, 21 f.

48 Townsend/Nielsen/Allister/Cassidy, Key ethical questions for research during the COVID-19 pandemic, The Lancet Psychiatry, 2020, 7 (5), 381 (382).

49 Hanna/Mwale, 'I'm Not *with* You, Yet I Am…': Virtual Face-to-Face Interviews/ Braun/Clarke/Gray, 2017, 256 (256 ff.); Hanna, Using internet technologies (such as Skype) as a research medium: a research note, Qualitative Research, 2012, 12(2), 239 (240); Iacono/Symonds/Brown, Skype as a Tool for Qualitative Research Interviews, Sociological Research Online, 2016, 21(12), 1 (4 f.).

50 Jowett, Carrying out qualitative research under lockdown – Practical and ethical considerations, Impact of Social Sciences, 2020, https://blogs.lse.ac.uk/impactofso

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

included in the study. The interviews were held in either English or German, depending on the participants' preference. These are the primary languages on the studied platforms and are suitable for conversations. However, the interviewers were sensitive to potential problems of lexicality and cultural misunderstandings.[51]

IV. Case studies of community-led platforms

Three case studies of community-led social media platforms with varying approaches to content governance are included after an overview of the desk research results about the platforms and their content moderation techniques.

1. mastodon

Mastodon is a federated microblogging platform where users can post messages, upload content, and communicate. Using the open protocol ActivityPub, it is part of the fediverse—the federated universe, which allows multiple websites and implementations to communicate with each other.[52] For the free and open-source project developed by Eugen Rochko, code, documentation, and policy statements are collaboratively developed.[53] Mastodon consists of multiple independent instances[54], each on its server, and each website can operate on its own. However, administrators can allow users to communicate with each other across websites. Users can either join existing instances (such as mastodon.social) or create their own using the mastodon software.[55] Some instances, such as mastodon.social, the flagship instance, limit the number of users to avoid

---

cialsciences/2020/04/20/carrying-out-qualitative-research-under-lockdown-practical-and-ethnic-al-considerations/.

51 Nakayama, Critical Intercultural Communication and the Digital Environment/Rings/Rasinger, 2020, 83 (87 ff.).

52 What is mastodon?, 2020, https://docs.joinmastodon.org/.

53 Zulli/Liu/Gehl, Rethinking the "social" in "social media": Insights into topology, abstraction, and scale on the Mastodon social network, New Media & Society, 2020, 22(7), 1188 (1189).

54 Instances are realisations of software running on a domain. For example, mastodon.social is an instance. See Instance, 2022, https://docs.joinmastodon.org/entities/instance/.

55 What is mastodon?, 2020, https://docs.joinmastodon.org/.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

centralisation.[56] Growth for a network is necessary, but it is preferred horizontally instead of within instances.[57] Once joined, users can see a timeline with posts of people they follow, and they can send short messages called toots, which they can boost, meaning share, and favourite, meaning to like.

Furthermore, users can make messages public for everyone, private for the user's followers, or direct to a specific mentioned user, and unlisted—visible to everyone but hidden in local timelines. Two timelines exist on mastodon: the local timeline, which only includes users' posts from the same instance, and a federated timeline, which provides for other instances if the user's instance is connected to them. That system creates a model similar to email.[58] That means users can send messages across instances[59] and the platform only allows that toots[60] are shown in chronological order, impeding any ranking due to algorithmic recommendations.[61]

Content governance is handled by each instance separately. Therefore, the content and posts that are allowed on the platform instances vary. Each instance can decide on permitted activities, whereby the most commonly prohibited activities are spam, nudity, and pornography.[62] The policy of each instance is available for users upon registration. For example, mastodon.social, a significant European instance with 7000 users as of December 2021[63], bans content such as Nazi symbolism and Holocaust de-

56 Leah & Rix, Bericht: Sparschwein, 2020, https://blog.chaos.social/2020/01/26/sparschwein-bericht-2020.html.

57 Zulli/Liu/Gehl, Rethinking the "social" in "social media": Insights into topology, abstraction, and scale on the Mastodon social network, New Media & Society 2020, 22(7), 1188 (1196).

58 Raman/Joglekar/De Cristofaro/Sastry/Tyson, Challenges in the Decentralised Web: The Mastodon Case, Proceedings of the Internet Measurement Conference, IMC '19: ACM Internet Measurement Conference, 2019, 217 (217).

59 Farokhmanesh, A beginner's guide to Mastodon, the hot new open-source Twitter clone, The Verge 2017, https://www.theverge.com/2017/4/7/15183128/mastodon-open-source-twitter-clone-how-to-use.

60 Toots are short messages with a limit of 500 characters. See Posting toots, 2020, https://docs.joinmastodon.org/user/posting/.

61 Zignani, Mastodon Content Warnings: Inappropriate Contents in a Microblogging Platform, Proceedings of the Thirteenth International AAAI Conference on Web and Social Media, 2019, 639 (640).

62 Raman/Joglekar/De Cristofaro/Sastry/Tyson, Challenges in the Decentralised Web: The Mastodon Case, Proceedings of the Internet Measurement Conference, IMC '19: ACM Internet Measurement Conference, 2019, 217 (220).

63 rixx, On Running a Mastodon Instance, 2021, https://rixx.de/blog/on-running-a-mastodon-instance/.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

nial, which is illegal in Germany. In addition, user accounts or content are deleted for posting illegal, discriminatory, violent, or nationalist content or content that harasses others. The instance not only makes rules public and transparent, but also explains best practices, for example, for crediting creators for art that is shared.[64] Even as each instance can implement their own rules, most instances use the standard ToS.[65] Instances can include users in content moderation to some extent as it includes a function for users to mark posts as inappropriate or sensitive.[66]

2. *diaspora**

The federated social network diaspora* consists of independently run servers called pods.[67] Users who register need to choose a pod, which allows crossposting with other services, such as Twitter. User data is stored in the chosen pod and each is managed by different people.[68] Community members are in charge of the pods and have access to the user data in their pods. Users can share and post content, including photos, videos, and music. They can mention others, like content, and interact with users from other pods.

Furthermore, they can follow people once they have set up their diaspora* identity. Moreover, diaspora* allows a function called aspects, which enables grouping users to and sharing content only with specific groups. Direct interaction with other users is only possible if they have been added to an aspect. Users are free to create their separate pods and change the source code to their liking.[69] Furthermore, users can share public posts accessible to all diaspora* users for sharing, commenting, and liking.[70]

Content governance depends on the administrators of the pods. Generally, removing pods or content is possible, as is flagging inappropriate

64 @ordnung, Rules, https://chaos.social/about/more.
65 mastodon, Datenschutzerklärung, https://mastodon.social/terms.
66 Zignani, Mastodon Content Warnings: Inappropriate Contents in a Micro-blogging Platform, Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019), 2019, 639 (639).
67 For a list of all available pods, see https://diaspora.fediverse.observer/list.
68 The DIASPORA* project, Choosing a pod, 2020, https://wiki.diasporafoundation.org/Choosing_a_pod.
69 diaspora*, Wie funktioniert diaspora*?, https://diasporafoundation.org /about#aspects.
70 The DIASPORA* project, FAQ for users, 2019, https://wiki.diasporafoundation.org/FAQ_for_users.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

user posts, which the platform recommends instead of replying. However, deletion is a last resort and conversations are a preferred remedy. Pods each have moderators who decide on the pod's rules. Moreover, users can flag inappropriate content and moderators can remove users and posts. However, new posts are not reviewed prior to publication.[71]

3. Slashdot

Slashdot is a website publishing written stories about various topics users submit. Editors check, format and correct the stories before publication, primarily including grammar, spelling, clarity, and link fixes. Furthermore, the editors "try to select the most interesting, timely, and relevant submissions".[72] Each user has an account with a nickname, which they use to comment on published stories. Users can filter the comments they see by choosing the range of a post's score (between -1 and 5), and they can report abusive comments such as spam or racism. Users can interact with each other in various ways, e.g., selecting someone as a friend makes the user their fan. If a user is selected as a foe, that respective user is called a freak.[73]

Slashdot involves users for content moderation. Users have karma (Terrible, Bad, Neutral, Positive, Good, and Excellent) depending on how their comments are moderated and the contributions they make to the site. Karma determines if users can moderate. Users get allocated moderation points they can spend on posts by rating them according to pre-selected categories, namely: Normal, Offtopic, Flamebait, Troll, Redundant, Insightful, Interesting, Informative, Funny, Overrated, and Underrated. Moreover, each comment has a score from -1 to +5, the default being at +1.[74] Posts are not deleted from the database; however, not all readers might be able to read them.[75]

Furthermore, Slashdot uses a method of meta-moderation to moderate the moderators. Therefore, users can also rate if moderators rated a story

---

71 The DIASPORA* project, FAQ for users, 2019, https://wiki.diasporafoundation.org/FAQ_for_users.

72 Slashdot, Frequently Asked Questions, 2022, https://slashdot.org/faq.

73 Slashdot, Frequently Asked Questions, 2022, https://slashdot.org/faq.

74 Slashdot, Frequently Asked Questions, 2022, https://slashdot.org/faq.

75 Lampe/Resnick, Slash(dot) and burn: distributed moderation in large online conversation space, Proceedings of the 2004 conference on Human factors in computing systems- CHI '04, 2004, 543 (544).

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

adequately and fairly. However, only users who have taken an active part in Slashdot for a longer time can metamoderate. Editors have unlimited moderation points, and they might ban IP addresses of users who exhibit abusive behaviour.[76]

V. *Advantages and disadvantages of content governance methods and the role of disinformation*

Community-led platforms employ various content moderation methods. Federated networks in particular exhibit significant variations in this regard. Table 2 summarises the advantages and disadvantages of the content governance models employed by mastodon, diaspora*, and Slashdot. It presents a summary of the collected material from interviews with administrators and moderators from the platforms.

76 Slashdot, Frequently Asked Questions, 2022, https://slashdot.org/faq.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

Table 2: Overview of advantages and disadvantages of content governance on diaspora, Slashdot, and mastodon.*

Approach	Advantages	Disadvantages	Used on
Deletion	Harmful content not accessible anymore	Might inspire trolls	m, d*
	Illegal content removed	Deletion of content that should be saved	
	No copycats	Harms freedom of expression	
	Could be reversed		
Deplatforming	Users can change their behaviour	Users might make a new account	/., m
	Temporary and reversible	Followers might continue	
	Enforce compliance		
	Removal of spammers/vandals		
Conversation	Clear misunderstandings	Admin resources	m, d*
	Open discussion pages	Incorrigible users	
	Enforce civility		
User moderation	Less admin resources	Too slow	/., d*
	Reduction of scale problem	Unfair users/bullies	
	Strengthen community		
Meta-Moderation	Moderator accountability	Biases by meta-moderators	/.
	Strengthens community	Remove admins' rights	
Hiding Content	Harmful content not accessible	Repost	d*
	Flexible and reversible		
Down-/Upvoting	Less visibility of harmful content	Organised trolls/bullies	/.
	Community decides		
Automated filters & human review	Reduces false positives	Psychological burden for moderators	m
	Easier to find harmful content		
	Reduction of workload		
Blocking	Protect users	Trolls would evade use	m
	Less workload for admins		

The most efficient methods to moderate problematic content are by no means universally endorsed. Generally, two types of governance can be

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

distinguished: one that is structured by the legal framework in the country the platform is operating in, and another that is structured by the algorithm of a platform (P6).

One interviewee, talking about diaspora*, argued that deletion is rarely effective at preventing abusive content, spam, or trolling (P10). Another one disagreed, arguing that deletion makes it more difficult for trolls to post offensive content (P3). However, deplatforming, meaning the removal of a user account due to infringement of platform rules[77], is difficult on decentralised networks because users could set a separate instance. Overall, however, it achieves the goal to limit the spread of undesired content, and even if users set up a separate node, larger nodes can significantly decrease the reach of the content by blocklisting them (P3). In other cases, users break the rules for understandable reasons and rules might need to be adapted (P2).

However, it is unclear whether deletion should imply the complete disappearance or marking of deleted content. Generally, deletion is sensible if a bot generates the content (P2). Furthermore, moderators are by no means perfect. They also have their agendas and sometimes do not act reasonably. Therefore, Slashdot developed a system of meta-moderators, leading to moderators' moderation. Most people moderate fairly, and the majority would not agree with unfair moderation. However, a system needs to be in place to filter out people at the extreme ends. It is not always easy to find enough people who are engaged and most platforms make profit from clicks. Love, porn, and lies keep people on the platform as users want to experience these emotions, so this content is favoured. To prevent that, Slashdot excludes users who spend too much time on the platform from moderation. In particular, if they post a lot about a specific topic, they cannot moderate that discussion (P14). Therefore, content moderation that is more objective is possible and reduces the spread of undesired content due to an increase of moderator accountability. However, resources and a sufficient number of administrators are necessary to realise that process.

Most community-led networks favoured communication with disruptive users to enforce civility. On mastodon, users are only removed if they are repeat offenders (P5). However, hiding content from all users except the author and moderators was mentioned as effective, especially against

77 Rogers, Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media, European Journal of Communication 2020, 35(3), 213 (214).

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

trolls, and banning users is helpful to decrease the administrators' workload (P10). Furthermore, on diaspora*, users can determine who interacts with them, delete comments about their posts, and define who has access. According to interviewees, even as users can do some content moderation, a moderation team with an overview is necessary to enforce community guidelines. However, finding the balance between moderator control and user self-sufficiency is challenging because too much control over user possibilities might lead to user experience regressions. In addition, users with insufficient technical expertise have difficulties exploring all their options. Some, however, are differentiated according to content if deletion is the best option. If the content can produce copycats or martyrs, especially in the case of violent acts, it should be deleted (P1).

Moreover, if disinformation causes direct harm to people, for instance, if they are added to pornographic content, deletion should be favoured. An example mentioned by the participant is Gamergate (P1). That so-called hashtag movement resulted in systematic harassment of women in the gaming industry by users who were reportedly frustrated with reporting about gaming.[78]

Giving users more control over with whom they can interact and over who can interact with them might help against harassment campaigns. That does not impede, however, the spread of disinformation. Nevertheless, controls that are more refined may also result in user experience regression (P3). For example, on mastodon, users can block other users or an entire instance. Furthermore, moderators can block instances and impede interaction with their instances (P5). On the instance mastodon.social, having an invitation-only instance has kept user interaction relatively civil.[79]

In some cases, the line between information and disinformation is blurred. Keeping up with current trends on disinformation is challenging for moderators and administrators as it takes a lot of time and can be psychologically draining. As one moderator on mastodon.social reports on moderation decisions:

> "We're under a lot of pressure to make the (or a) right decision, to prove we deserve the trust people generously placed in us. More than

78 Massanari, #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures, New Media & Society 2017, 19(3), 329 (330).

79 rixx, On Running a Mastodon Instance, 2021, https://rixx.de/blog/on-running-a-mastodon-instance/.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

> that: We have to make a decision that we're still willing to live with then it's thrown back at us by a malicious troll twisting the facts, while we can't respond properly without disclosing messages we're committed to protect. You start second-guessing every word and try to see what everything would look like when it's taken out of context. It's a defensive and indefensible position."[80]

Automated content moderation would be tricky, as AI systems cannot consider contextual aspects within a culture, and they cannot read between the lines (P1, P9) or recognise nuance (P8). However, automated content moderation can support moderators and ease their workload (P2), for instance, by removing comparatively unambiguous material such as spam (P4, P8) or flagging posts that might distribute undesired content (P8, P10). Still, automated systems might be subject to abuse and have a high chance of resulting in false positives (P3). Furthermore, detecting and reducing spam and content by trolls is an arms race because they keep posting and often use automated means themselves (P9, P14). However, deletion of such content is still essential for allowing the communities to continue functioning (P10), and sometimes "starving the troll", meaning to deprive them of the attention they seek, is the best option (P10). Human moderators, however, would need education and training to make the right decisions, and on many platforms, not enough moderators are working to cover all the content produced (P7). Moreover, automated systems as decision-support might bias moderators, so they might be less prone to question the AI's decision to remove content.[81] They might enforce existing taboos, strengthen echo chambers, and drive people into smaller, invisible communities (P10). Some automated frameworks have been developed to detect disinformation[82]; however, once identified, how to react to the content is not entirely clear, as solutions need to be adapted to the contexts of different platforms.

---

80 rixx, On Running a Mastodon Instance, 2021,https://rixx.de/blog/on-running-a-mastodon-instance/.

81 Gsenger/Strle, Trust, Automation Bias and Aversion: Algorithmic Decision-Making in the Context of Credit Scoring, Interdisciplinary Description of Complex Systems 2021, 19(4), 542 (547).

82 Kim/Tabibian/Ach/Schölkopf/Gomez-Rodriguez, Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation, WSDM '18: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, 324 (324); Alaphilippe/Gizikis/Hanot, Automated tackling of disinformation, European Parliament, 2018, 1 (35 ff.).

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

Transparency and balance of content management practices are crucial to keeping the users' trust, which is based on the moderators' intuitions to decide what might hurt but still be permissible content (P1). In that regard, deleting user-profiles and instances is the last resort and, most of the time, only happens to repeat offenders.[83] Therefore, automated systems should not make the final decisions, as they are rarely held accountable for their actions (P10). Granting users the ability to manage their spaces might be a solution (P2), and in many community-led platforms, that is the case. Moreover, these platforms do not employ any algorithm-based content distribution, which might decrease the distribution of high-affect content, such as disinformation.[84]

VI. Conclusion

The cases of community-led platforms discussed in this article demonstrate the potential diversity of content governance models. In contrast to the responses to problematic content of very large online platforms, which have become increasingly stratified around content deletion and moderation, community-led platforms' responses to problematic content are much more diverse. The diversity and richness of their responses demonstrate what is possible and how much could still be done to expand existing understandings of what could reasonably constitute content moderation.

At the same time, community-led platforms also have a different set of incentives and typically operate at a different scale than very large online platforms. This raises the open question of whether the things that can be learned from community-led platforms can be applied beyond these platforms.

We believe that there is indeed space for large platforms to learn from smaller community-led ones. The innovative techniques applied by smaller platforms are not significantly limited in scope or scale. Still, they do require additional effort and investment as well as breaking open stratified ways of thinking about content moderation. Here regulatory interventions also have a role in ensuring that they create the business incentives and the regulatory environment that enable this kind of design innovation.

---

83 A list of all blocked instances can be found at https://github.com/chaossocial/about/blob/master/blocked_instances.md.

84 Acerbi, Cognitive attraction and online misinformation, Palgrave Communications 15, 2019, 1 (2 ff.).

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb

Notably, both the current draft of the EU Digital Services Act[85] and the UK Online Harms Bill[86] have integrated language into them that attempts to influence platform design and encourage a race to the top rather than a struggle to the bottom. Whether they are successful in creating incentives to improve the design of large online platforms remains to be seen.

Beyond the potential influence of community-led platforms on larger online platforms, their creative approaches to the challenges of problematic online content remain a valuable perspective for re-imagining how content moderation takes place at present. Instead of seeing the troubling business practices of large online platforms as inevitable, the public debate needs to embrace the full diversity and potential of existing content moderation practices. We hope that this article can provide a contribution to expanding the imagination of what is possible in terms of content management.

85 European Commission/Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC.

86 House of Lords/House of Commons/Joint Committee on the Draft Online Safety Bill, Report of Session 2021–22.

https://doi.org/10.5771/9783748934981-181, am 30.11.2022, 17:11:12
Open Access - (cc) BY-SA - http://www.nomos-elibrary.de/agb