REIMAGINING CONTENT MODERATION AND SAFEGUARDING FUNDAMENTAL RIGHTS

A STUDY ON COMMUNITY-LED PLATFORMS

Ben Wagner, Johanne Kübler, Eliška Pírková,
Rita Gsenger, Carolina Ferro

May 2021

THEGREENS/EFA
in the European Parliament

Acknowledgements

This study was commissioned by the Greens/EFA group in the European Parliament. The authors wish to thank the European Greens/EFA team, especially Kirsten Fiedler, Narmine Abou Bakari, and Eszter Bako, for their fantastic support. We would also like to acknowledge the valuable contribution of the people we interviewed during the process of elaboration of this study, without whom this study would not have been possible. Thank you for dedicating your time to reply to our questions, even in such uncertain and difficult times.

Table of contents

Introduction

As millions of people log into social media platforms like Facebook, Twitter, and Reddit every day, the rules by which these platforms are governed increasingly determine how we interact with each other, and they shape the possibilities and nature of public discourse (Suzor 2020). The kinds of rules these platforms adopt, the policies they choose to enact, and the design choices they make all affect which information is available and how people communicate.

In line with the steep increase in user numbers and data generated per day, technology companies had to develop procedures to process a previously inconceivable amount of data. For instance, Facebook generates an estimated 4 petabytes of data every single day (Osman 2021). Responding to the need to examine and curate a large amount of data, platforms have developed content governance models and complex, multi-layered content moderation systems, which rely heavily on the removal of harmful and, otherwise, undesirable content.

However, there are growing concerns regarding the impact of those platforms' decisions on the freedom of expression and information and the digital rights of individuals. The focus on the blocking and deletion of content is accentuated through legislative approaches that also focus on deletion. In recent years, increased governmental pressure on online platforms "to do more" about the spread of hate speech, disinformation, and other societal phenomena online has led to a frenetic regulatory process across the European Union (EU), which, consequently, triggered similar regulatory responses around the globe. Due to the lack of legal certainty in combination with unduly short time frames for content removal and the threat of heavy fines for non-compliance, platforms frequently over comply with these demands and swiftly remove large amounts of online content with no transparency and public scrutiny (Dara 2011; Ahlert, Marsden, and Yung 2004; Leyden 2004). The sheer volume of requests inevitably leads to erroneous takedowns, resulting in chilling effects for users faced with them (Penney 2019; Matias et al. 2020).

Many community-led platforms[1] offer alternatives to these challenges for human rights and freedom of expression. However, these innovative approaches are typically not implemented by larger

[1] Community-led platforms are platforms partially or entirely governed by its community of users.

platforms. The alternative approaches often focus on community building and curation to strengthen communities to the point that content moderation is considerably less necessary. To accurately assess these alternative approaches, it is important to closely analyse the effects of different types of content moderation on user behaviour and their digital rights. Online communities without any content moderation at all are equally problematic for digital rights and typically quickly descend into what Daphne Keller termed the freedom of expression 'mosh pit.' (The Verge, 2021) Such online communities are not ideal for any of the actors involved, as only the loudest actors' voices can be heard.

This study explores alternative approaches to content moderation and, overall, different content governance models. Based on the research outcomes, it provides a set of recommendations for community-based and user-centric content moderation models that meet the criteria of meaningful transparency and are in line with international human rights frameworks. These recommendations are specifically addressed to EU lawmakers with the goal of informing the ongoing debate on the proposed EU Digital Services Act.

Content governance theoretical framework and methodological approach

Different content governance models are used by online platforms that heavily rely on user-generated content, such as social media platforms, online marketplaces, communities and forums. These online platforms abound in content created and shared organically by users, such as text, images, video, posts, tweets, and reviews.

In this study, 'content governance' refers to a set of processes, procedures, and systems that define how platforms plan, publish, moderate, and curate content. 'Content moderation' is the organised practice of a social media platform of screening content to guarantee its compliance with laws and regulations, community guidelines, and user agreements, as well as norms of appropriateness for that site and its cultural context (Roberts 2017). Content moderation practices can be distinguished as: i) the removal of illegal content; ii) content in violation of Terms of Service (ToS); and iii) content deemed "unacceptable", "disruptive" or "inappropriate" by consensus of the community. Each type will be outlined below.

Many countries oblige social networks to remove any content that is "manifestly unlawful". EU law outlaws four types of content: (i) child sexual abuse material; (ii) racist and xenophobic hate speech; (iii) terrorist content; and (iv) content infringing intellectual property rights. Beyond these categories, what is considered illegal content varies widely among member states. Thus, "the same type of content may be considered illegal, legal but harmful or legal and not harmful" across EU member states (De Streel et al. 2020). As a result of the ambiguities in law, internet companies developed and instituted their own content moderation practices.

Beyond the manifestly illegal, social networks remove content in contravention of their own ToS (also known as Terms of Use or Terms and Conditions). ToS are a legal document a person must agree to abide by when registering an account. ToS governs many areas, from users' control over their privacy to disclosures on how a given platform is allowed to use their data. However, ToS are usually phrased in a broad and vague manner, giving platforms considerable discretion in what content they actually choose to act on. Furthermore, certain groups can be excluded from using a service, for example, children under the age of 13.

Lastly, community guidelines direct the behaviour of all community members during their participation in that community, setting out which content and behaviours are deemed “unacceptable”, “disruptive“ or “inappropriate”. These guidelines typically are constantly evolving. Community guidelines can arise from a consultation of the online community during which they are developed and refined, but many are conceived in-house by social media companies. In addition to public community guidelines, online platforms maintain an internal document. This internal document, a much more in-depth version of community standards, guides their human moderators. Thus, while some communities stress a community consensus to determine which behaviours are acceptable and which are not, the main tools for content moderation are the ToS, community guidelines, and internal guidelines for moderators.

Taking into consideration the possibility of adopting different content governance models that apply varying content moderation practices, this study aims to explore the following questions:

- Which alternative content governance models (and the content moderation practices within them) exist among community-led platforms?
- What are the advantages and drawbacks of identified alternative content governance models used by community-led platforms?
- Is there any empirical data that allows us to measure the effectiveness of the alternative content governance models used by community-led platforms?
- Which approaches have the potential to be more widely employed by major social media platforms?

I. Theoretical framework

Participants in online communities often have different and sometimes competing interests. Over time, a rough consensus about the range of behaviours deemed acceptable and others considered unacceptable emerges among members of a given community. The kinds of behaviours viewed as acceptable, or normative, and those which are not, vary. For instance, contributors to Wikipedia are required to adopt a neutral point of view, and appraisals in content recommendation systems, such as TripAdvisor and Yelp, are expected to be genuine reviews of customers of a particular venue. A rough consensus about acceptable behaviours within a community can help it achieve its goals. As such, the neutral point-of-view norm in Wikipedia furthers the community’s goal of writing a trustworthy encyclopaedia (Wikipedia 2020a). In many technical support communities, responses to questions are expected to be supportive, rather than antagonistic, of furthering their mission of providing members with assistance to deal with problems they are facing.

Because conflicts are an inevitable feature of social interactions, they are also a frequent occurrence in online communities. Online communities face many challenges from external actors as well as internally. External agents may either aim to disrupt a given community or manipulate it for their gain. Outsiders posing a threat to communities include so-called trolls—that is, people posing as legitimate members who post controversial, inflammatory, irrelevant, or off-topic messages designed to provoke others into an emotional response (Schwartz 2008)—and manipulators who seek to influence the community to produce certain outcomes, such as more favourable reviews for a commercial establishment. As outsiders, trolls and manipulators lack a vested interest in the community's health. This makes it particularly difficult to deal with them because social sanctions, which are likely to be effective when dealing with internal challengers, risk having either no effect or being counterproductive by increasing their activity (Kiesler et al. 2012:127).

Insiders—that is, members invested in the community who care about the community's health and their standing in it (Kiesler et al. 2012:140)—may also breach behavioural norms. For instance, insiders may violate a community's norms due to ignorance, because they fail to infer rules from observation, or because they contest existing behavioural norms (Kiesler et al. 2012:129).

Frequent breaches of behavioural norms and protracted conflicts can inflict serious damage to online communities. Thus, it is desirable to decrease the frequency of non-normative behaviours or lessen their impact on the community. This study argues that online communities can be designed to attain societal or public interest goals and not for economic gain. In fact, simple design, communication, and framing choices can significantly impact norm compliance in online communities.

In his landmark study about law and cyberspace, Lessig (1999) identified four elements that regulate behaviour online: law, norms, market, and architecture (or technology). Given that online community designers' control over the laws governing their communities is limited, Sara Kiesler et al. (2012) proposed design alternatives that can be grouped into the following three categories: norms, market, and technology. For example, the adoption of certain design decisions can make norms more salient, and greater compliance can be achieved. Economic incentives include reputation systems and internal currencies. Lastly, moderation systems and reversion tools are technical ways to prevent bad behaviour and restore previous versions of the content. Oftentimes, these measures work most effectively in combination.

The theoretical framework that follows will be based primarily on *Regulating Behavior in Online Communities* (Kiesler et al. 2012), *Governing Internet Expression* (Wagner 2016), *Custodians of the Internet (Gillespie 2018), Content or Context Moderation* (Caplan 2018), *Behind the Screen* (Roberts 2019), and *Content Moderation Remedies* (Goldman 2021). The theoretical framework will assist in understanding different content governance models and approaches to online content moderation. First, we will present how the damage caused by bad behaviour can be reduced, followed by ways to control the amount of bad behaviour that any bad actor can engage in. Second, we will explore ways to encourage compliance with norms through psychological and economic incentives.

Limiting the effects of bad behaviour

Content moderation—that is, the process of pre-screening, labelling, moving, or removing inappropriate content—is a widespread practice in the industry. Thus, the damage such messages can cause is limited because the number of people who will read them is reduced.

Within the online community, content moderation can be perceived as controversial when the moderator's decisions are perceived as illegitimate, potentially leading to more disruption (Gillespie 2018). Thus, it is important to increase the acceptance of moderation actions, for example, by refraining from silencing the speaker and redirecting them to other places instead of removing inappropriate posts (Kiesler et al. 2012). The legitimacy of the moderation process and thus the effectiveness of moderation decisions can be further increased by a consistent application of moderation criteria and by providing a chance to argue one's case, as well as appeal procedures (Jhaver et al. 2019b). Also, moderation is perceived as more legitimate and, therefore, more effective when done by people who are members of the community, who are impartial, and who have limited or rotating power.

In production communities—that is, communities whose members jointly produce a good—reversion tools have proven useful to limit the damage done accidentally or through vandalism, for example, for Wikipedia entries. In content recommendation systems prone to manipulations, the equivalent to moderation and reversion is to use algorithms that look for suspicious patterns to filter out or discount suspicious ratings (Caplan 2018). Given that trolls feed on the community's reaction to their inflammatory and provocative actions, establishing norms for ignoring trolls greatly limits the damage they can inflict on a community.

Coerced compliance: Limiting bad behaviour

Individual bad behaviour usually causes only limited harm. However, repeated disruption can cause severe damage and thus needs to be curtailed. For instance, repetitive spam-like activity can lead to large-scale damage in an online community (Blackwell et al. 2017). Many platforms have developed *throttles* or *quota mechanisms* that block or send warnings to users who post too many messages in too short a time, for example. These activity quotas allow people to participate in a community but prevent repetitive, spam-like activity (Kiesler et al. 2012). Another common practice in online communities is to limit the continuing damage of offenders with gags or bans. Gags and bans are only useful if it is difficult for the bad actor to use a different account or if the ban is disguised (Goldman 2021). As with moderation decisions, acceptance and thus the effectiveness of bans and gags increases when criteria are applied consistently, members are able to argue their case, and appeal procedures exist (Jhaver et al. 2019b).

Some communities choose to limit damage by requiring members to earn the privilege of taking certain potentially harmful actions, for example, through internal currencies or ladders of access. Internal currencies are accumulated easily through everyday participation, such as providing truthful information, but they are difficult to acquire by trolls and manipulators, thus limiting their influence. Lastly, the creation of fake accounts for automated attacks by spammers and other strategies can be limited with CAPTCHAs[2] or identity checks before validating subscription.

Encouraging voluntary compliance

While damage inflicted by bad actors like trolls and manipulators can be limited and compliance with behavioural norms can be coerced, there are techniques for achieving voluntary compliance with behavioural norms. Insiders, who care about the community's health and their own standing within

[2] A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a test presented to a user that should be easy for a human to pass but very difficult for a computer.

are more susceptible to respond favourably to measures to achieve voluntary compliance.

To be able to follow the rules and norms, community members must be aware of them. People tend to infer behavioural norms from:

1) Observing other actors and the consequences of their behaviour;
2) Seeing instructive generalisations or rules of conduct;
3) Behaving and directly receiving feedback.

Given the propensity of people to learn from examples, platform creators and managers enjoy considerable power when it comes to framing, which is through design choices, communicating which behaviours are normative. One measure is to make norms clear and salient, for instance, by prominently highlighting examples of desired behaviour and contrasting examples of inappropriate behaviour (Leader Maynard and Benesch 2016). While increasing the awareness of members to the community's norms, prominently displayed behaviour guidelines may signal to them that guidelines are not always followed (Kiesler et al. 2012). One way to counter this impression is to make rules and guidelines prominent but only at the point where people may be about to violate them.

While awareness of norms is a first step towards adherence to them, it does not guarantee compliance. Compliance can be increased by promoting cohesion of the community, and including the community in the crafting of the rules makes them more legitimate and members more likely to conform to them. Furthermore, people are more receptive to discontinuing bad behaviour and change when disciplinary actions are presented in a way that allows people to save face ("Someone using your account…," or "You may not have realised…"). Compliance can be further increased by measures such as reputation systems, such as one that condenses the history of a member's online behaviour into a score (Goldman 2021). Lastly, authentication of identities may discourage norm violations, as does incentivising users to maintain a long-term identifier in communities relying on pseudonyms (Caplan 2018).

When facing disruptive behaviour, creators and managers of online communities might be inclined to resort to a tangible solution, such as removing inappropriate posts, banning or throttling the posters (Wagner 2016). While damage created by outside actors such as spammers and manipulators who operate many accounts and employ bots may need to be limited through automated and tangible measures, less tangible, behavioural approaches provide a useful first effort to address harmful behaviour by community members.

Making norms clear and salient and building legitimacy are effective ways to increase voluntary compliance (Caplan 2018). Individuals engaging in bad behaviours can be corrected in ways that allow them to save face, and trolls can be ignored. Although behavioural responses may not be sufficient to address the whole range of bad behaviours in online communities, they represent a fantastic toolbox to promote normative behaviours and thus help online communities to thrive.

II. Research design, methodology and case selection

Prevalent content governance models rely their content moderation primarily on the identification and subsequent deletion of unacceptable content, with a significant impact on human rights online such as freedom of expression of their users. To explore alternatives to these flawed practices, this study explores community-led alternative content governance models. These community-led platforms are unusual within the industry (Mills, Durepos, and Wiebe 2010). However, their alternative practices with regard to content moderation make them crucial to assess which of their approaches show potential for the industry at large. We explore how relatively similar community-led platforms employ distinct strategies regarding content governance, with a focus on community building and curation, thereby strengthening communities to make intervention through content moderation less likely.

This study relies on a qualitative research design. A qualitative research methodology is based on a descriptive narrative for data analysis. It is appropriate for studies that encompass relationships between individuals, individuals and their environments, and motives that drive individual behaviour and action (Berríos and Lucca 2006). Qualitative methodology and its research methods focus on the whole rather than the parts (Gerdes and Conn 2001). Therefore, this methodology allows us to adequately address this study's research questions and, hence, is most appropriate for investigating community-led alternative content governance models.

We employ qualitative research methods, notably "case studies", to investigate community-led online platforms applying alternative content governance models. A case study is an "empirical method that investigates a contemporary phenomenon (the "case") in depth and within its real-world context, especially when the boundaries between phenomenon and context may not be clearly evident....[and] relies on multiple sources of evidence" (Yin 2018:45), allowing the study of complex social phenomena. The case studies will be examined for commonalities; thus, the rationale behind the selection of the cases is a most similar case design. The most similar systems design stems from the idea that "all cases share basic characteristics but vary with respect to some key explanatory factor" (Moses and Knutsen 2007:98). Multiple methods were used to gather data for the case studies: **semi-structured interviews** (with moderators and administrators of different online platforms and researchers on the topic) and a **literature review** of existing publications on the topic.

In addition to data relating to the case studies, a literature review and interviews were conducted with actors related to other community-led online platforms and academics researching the topic. Whereas the case studies showcase some aspects of alternative content governance models, the additional material serves to discuss other alternatives and different perspectives on alternative content governance models.

Case selection

This study will present five case studies of community-led social media platforms employing alternative approaches to content governance and moderation. Each case study illustrates a distinct approach, as follows:

- **Wikipedia:** The open-collaborative online encyclopaedia is an example of the use of technology to promote a platform's mission. The website resorts to automated moderation through bots, who carry out repetitive and mundane tasks to maintain the 52 million pages of English Wikipedia and promote good behaviour among its community of volunteer editors (Wikipedia 2021g).

- **diaspora*:** The group of independently owned servers forming a federated social network is an example of a norms-based approach. The individual nodes of the federated network enjoy a high degree of autonomy to the point that the core team moderators do not have access to all discussions or other nodes, leading to the creation of a separate discussion channel (diaspora* n.d.b).

- **Mastodon:** The decentralised, open-source, and increasingly popular Twitter alternative is an example of a platform based on a pluralism of rules. Each community chooses its own rules and enforces them through moderators. Contrary to Reddit, where subreddit communities are still governed by Reddit administrators whose control spans across the site, Mastodon communities are structurally independent of one another.

- **DER STANDARD:** The website of one of the leading Austrian daily newspapers experimented with a norm-based approach to decrease the amount of hate speech in its comment section. The experiment showed users videos of the real-life consequences of posting hate speech in its discussion forum and was considered a success.

- **Slashdot:** The social news website is a good example of a ranking and rating-based system. It features a community-based moderation system in which members rate the quality of contributions to encourage contributions that meet the community's norms of quality.

Semi-structured interviews

For this study, 16 semi-structured interviews were conducted with moderators, administrators, various online platforms, and researchers to discuss alternative content moderation practices and content governance models. Qualitative interviews are a powerful tool to explore how the respondents understand their environment, actions, and roles in a given online platform setting. Semi-structured interviews are particularly useful for understanding the sense that actors give their actions. Therefore, it is essential that their testimony is obtained in an unobtrusive, non-directed manner (McCracken 1988:21).

Conducting qualitative research, including interviews, during the Covid-19 pandemic is beset with challenges. Notably, social distancing measures and travel restrictions were in place at various moments in 2020 and 2021. Furthermore, face-to-face interaction for data collection through interviews potentially puts respondents and interviewers at risk of contracting Covid-19 and must, therefore, be deemed unsafe, unethical, and even illegal (Townsend et al. 2020).

The most suitable alternative to face-to-face interviews is the use of video-calling to generate the required data (Braun, Clarke, and Gray 2017; Hanna 2012; Lo Iacono, Symonds, and Brown 2016). However, this method is not entirely without challenges, such as participants not being able to use the technology or having a poor Wi-Fi connection (Jowett 2020). Furthermore, some study participants face time constraints due to additional childcare obligations and are therefore not available for video-calling. Given these difficulties, this study combines written interviews and qualitative video-calling conversations using a semi-structured questionnaire (see Annexes 6.1-6.3).

The semi-structured interview format consisted of a flexible list of questions that were applied situationally (Kruse 2014). The interview grid contained basic information, such as the interviewee's name, affiliation and number of active years on the studied platform, situating the interviewee in a wider context and enabling us to better understand their responses. The grid also contained a list of open-ended substantive questions. These very general questions, phrased in a nondirective manner, were intended to guide the conversation between the interviewer and the interviewee while allowing respondents to tell their story in their own terms (McCracken 1988:34).

The interviews were conducted in either English or German. These are the primary languages on the studied platforms and thus lend themselves well to data collection. However, the interviewer is sensitive to potential problems of lexicality and intercultural (mis)understanding for participants whose first language is not English or German (Rings and Rasinger 2020).

Overview of the cases

For this study, five platforms applying alternative content governance models were investigated in more detail: Wikipedia, diaspora*, Mastodon, Der Standard, and Slashdot. The following sections provide an overview of how these platforms work and a summary of their content governance models. These include results from the qualitative interviews with content moderators, including studies concerning the functioning and challenges of their models of content governance.

I. Wikipedia

Wikipedia is a free online encyclopaedia supported by the Wikimedia Foundation. Volunteers provide the encyclopaedia's content. Currently, Wikipedia comprises 55 million articles in 300 languages (Wikipedia 2021a). The site operates according to five fundamental principles or pillars: Wikipedia is an encyclopaedia; it is written impartially; it consists of free content anyone can use, edit, and distribute; its editors should treat each other respectfully; and it does not have any firm rules, only guidelines (Wikipedia 2021d).

Content governance on Wikipedia revolves around a self-regulatory community of voluntary administrators and editors who discuss any changes or additions in content collectively and publicly. Content is reviewed either before it is published or after, depending on the rules of the different Wikipedia communities. Depending on their standing in the community, moderators have varying rights and possibilities to decide about the content on the sites.

The different language versions of Wikipedia are based on the same idea but have some differing features. The software is the same, but each language version uses differing extensions and applies different rules for content moderation. For instance, in German Wikipedia, each new edit needs to be approved by established authors, which is not the case for the English Wikipedia. Moreover, the authors' locations differ, as English Wikipedia's members are distributed globally, whereas the majority of German Wikipedia's contributors are located in Germany. Therefore, each language version has its own work culture, quality standards, and etiquette (Merz 2019). The English, Swedish, and German

Wikipedias are the largest active communities as of February 2021, measured by the number of articles published (Wikimedia 2021). Lately, the different language versions of Wikipedia are influenced by the Wikimedia foundation's software on which it runs.

Wikipedia does not have moderators like social networking sites, but it has administrators or sysops (system operators).[3] They are editors who are granted certain abilities, for instance, to block and unblock user accounts, apply or change page protection, delete, undelete, and rename pages. Any user can register to become a voluntary editor on Wikipedia and write or edit pages. Administrators are volunteers as well and are usually granted that function after a community review process. Any registered user can request to become an administrator, although only users with considerable experience are granted that status. For a duration of seven days, editors ask the candidate questions, and a discussion takes place until a user can become an administrator.

Administrators can help solve disputes, try to decrease vandalism, and enforce sanctions. They should do so impartially, behave civilly, avoid personal attacks, and act in good faith. If administrators repeatedly violate these principles or lose the community's trust, they might have their admin rights removed by an Arbitration Committee (Wikipedia 2021b). Administrators and editors can take several actions if the content is deemed harmful or controversial. The deletion of content depends on the context, whereby some types of content, such as illegal material, need to be removed. For other types of content, other solutions might be more viable.

Due to the complexity of governance in Wikipedia, the barriers to entry are higher. New users might not know what they are allowed to write. As long as editing and writing follow guidelines, they can be interpreted and changed flexibly. As soon as the guidelines are translated into rules enforced by software, they cannot be adapted so easily. Therefore, if specific words are not permitted, an article is not published, and the user might not know why. In that context, the question is which rules should be translated into software, because as soon as it is written down, that rule cannot be circumvented or questioned that easily. Moreover, some rules and guidelines might change, which is more difficult to adapt at the software level.

Generally, anyone can contribute to Wikipedia and edit articles, either as a registered user or anonymously. In the latter case, the edits are attributed to the IP address of the user. However, in some cases, editing Wikipedia articles is limited due to vandalism (Wikipedia 2021a). Blocking users is the last resort against vandalism. Users who spam and vandalise actively and persistently can be reported (Wikipedia 2021c). English Wikipedia is often facing cases of contributions that are considered to be vandalism. Due to many instances of vandalism, Wikipedia administrators have developed several tools to detect such edits or articles. For instance, an edits' list of each page can be accessed and filtered according to the type of entry, content, or editors, such as anonymous or new editors.

If vandalism is detected, administrators can take several steps, such as blocking, protecting or deleting pages. Moreover, the page could be reversed to its original state before the occurrence of vandalism,

[3] The following only reflects the rules and conventions for the English Wikipedia. For other languages, the guidelines and rules deviate.

even in cases of multiple edits. Tools were developed to increase the efficiency of these measures, such as edit filters, which might prevent an edit from being posted. Moreover, algorithms are used to detect vandalism in edits. These include, for instance, the language used, other textual features, the metadata of edits, and the reputation of the editor. Furthermore, algorithms assign a score to the edits made and list them according to the probability of being vandalised. A human can subsequently evaluate whether the edits selected by the algorithm need any intervention. Several bots are active in scanning any incoming edits. If the bot identifies an edit to be vandalism, it reverts the edit and leaves a note on the vandal's talk page. Some of these bots operate on neural networks, continuously learning to distinguish between good and bad edits (De Laat 2015).

Aside from vandalism, Wikipedia needs to cope with personal attacks on talk pages. English Wikipedia's guideline prohibits personal attacks on editors, for instance, abusive, discriminatory, and defamatory language or threats (Wikipedia 2020b). A 2017 study about English Wikipedia concluded that anonymous contributions were more likely to be an attack (Wulczyn et al. 2017). Less than half of the attacks are were by anonymous users. Moreover, the majority of attacks were initiated by users who were not very active on the platform. Overall, 13% of attacks prompted a response by human moderators within 7 days. Therefore, the paper concludes that automated classifiers might be valuable for moderators on Wikipedia. Due to the large amount of content and activity on Wikipedia, human moderators alone cannot review every edit and contribution. Therefore, a combination of algorithmic and human moderation makes sense to Wikipedia, even though both have their flaws and might be biased.

II. diaspora*

The diaspora* project is a decentralised, federated social network existing on independently run servers. As of January 2021, 101 pods with sign-ups are registered in various countries, such as France, Germany, Hungary, Switzerland, the United States, and Canada (Poduptime, 2021). Diaspora* runs on free software in which any user can change the source code to improve the network and contribute to its maintenance (diaspora* n.d.a). The community maintains an official wiki to provide information, documentation, and developer resources (The diaspora* project, n.d.). Moreover, a discussion page about various topics, including announcements by the diaspora* core team, development, support, and project management, is sustained (diaspora* n.d.b).

Content governance on diaspora* is dependent on the administrators of the pods who set up or are in charge of a network. Therefore, no common guidelines can be inferred. Overall, removal of content and entire pods is possible, as well as flagging posts by users. No content is previewed, deletion is generally regarded as a last resort, and conversations are preferred.

The network runs on independent servers, so-called *pods*. Therefore, it does not have a central base or hub. Users can register with a pod of their liking, and their data remain only with that pod. The administrators of the pods a user interacts with solely have access to users' personal data, which are only used to allow functions of the network, such as connecting with others and sharing content. Moreover, each user can set up and host their own pod. Communication between pods is easily possible, and users can interact with their contacts across the network. Each user creates a diaspora* identity (ID) as username@podname.com, which does not need to correspond with their real identity.

The diaspora* network enables different forms of interaction, such as following people, sharing photos, videos, music, and other content. Users can add others with whom they interact to different *aspects* (diaspora* n.d.a). Aspects are a way of grouping user contacts on the network and sharing only certain chosen parts of a user's profile and content. Sharing is asymmetrical; so even if a user shares with one of their aspects, the other users in that aspect might not be sharing anything. Additionally, a user can share with another user only by adding them to an aspect. Users who are not added to any aspect can see only public posts by that user. Public posts can be viewed, commented on, and shared by any logged-in diaspora* users (The DIASPORA* project, n.d.a). Overall, aspects allow control over the content users share and with whom they share it. Users can share content only with certain parts of their network—some aspects—or with the entire network. Diaspora* allows for similar features as other social networks, such as hashtags, resharing, mentioning other users, and loving content (diaspora*, n.d.a).

Each pod has its own moderator(s) who set up their own rules for that part of the network. Moderators are responsible for the pod, and each user can flag problematic posts. Moreover, flagging posts is recommended instead of replying to problematic posts. If enough users flag a post, the moderators will take action. Generally, new posts are not previewed. Moderators have the right to remove posts or users anytime (diaspora*, n.d.b). However, most diaspora* communities are rather small, so automatically deleting content is not feasible.

According to one interviewee, deletion is rarely effective in preventing intentionally abusive content, automated spam, manual spam, and trolling. Another disagreed, as they argued that deletion makes it annoying for trolls to post content, as they would not get the attention they would want. Deplatforming—the removal of a user account from a platform due to infringement of platform rules (Rogers 2020)—is especially difficult in a decentralised network such as diaspora*, as individuals could just set up their own instance.

Furthermore, moderators would rather communicate with offenders to enforce civility. Instead of deletion, content is hidden to all pod members, except the author(s) and moderator(s). To keep a community healthy and the moderators' workload manageable, administrators might ban users. At diaspora*, some moderation tasks are entrusted to the users. They can determine who interacts with them and their content. More specifically, users can delete comments on their own posts or ban other users from interacting with them. Moreover, users can limit the range of people with access to their posts. This is effective, especially as the diaspora* community does not define a specific topic and does not have common global community guidelines.

Sometimes, however, the delegation of moderation to users can cause disputes and schisms because diverse viewpoints are less likely to be discussed and confronted. A global, more objective moderation team enforcing community guidelines is still needed, even as users can moderate a little. This kind of user control is especially effective against targeted harassment campaigns. On the downside, finer controls for users might result in user experience regressions, whereby users with less technical experience have difficulties handling options and their consequences.

III. Mastodon

Mastodon is a federated social media platform comprising various nodes, including many different topics where users can post messages, images, and videos (Farokhmanesh 2017). Mastodon is a free and open-source project with a general public licence in which code, documentation, and policy statements are collaboratively developed (Zulli et al. 2020).

Content governance on Mastodon is dependent on the rules of the different instances of the federated network, similar to diaspora*. Central to the success of Mastodon's content moderation is the limitation of user numbers. Furthermore, content warnings are used, and conversation is preferred to deletion, the latter being used only as a last resort.

The flagship instance is called mastodon.social, with currently 7000 accounts and 2500 weekly active users. Overall, the instance has nearly 2 million statuses. Currently, mastodon.social does not accept new users to avoid centralisation of the network (Leah & Rix, 2020). For administrators and users, the number of instances and the level of engagement in each instance are more important than drawing a large number of users. A certain number of users is necessary for a social network to function; however, growth is emphasised horizontally instead of within instances. Furthermore, a limited number of users is more likely to maintain the quality of engagement (Zulli et al. 2020).

Many other instances exist that vary regarding topics and rules concerning posts they allow. The network looks similar to Twitter, as it shows users a timeline of people they follow and users can send short messages with up to 500 characters called *toots*, users can *favourite,* that is, like toots and they can share toots, which is called a *boost*. The language can, however, vary according to the specific instances. Mastodon offers several options for the privacy of messages: public, private, unlisted and direct. Public messages are available to everyone in the local timeline, private is for the sender's followers, and direct messages go to the user mentioned in it (Farokhmanesh 2017). Unlisted toots are messages visible to everybody but not included in the local timeline (mastodon, n.d.) Moreover, users can hide content behind content warnings if they might be considered inappropriate for the community. For instance, users can make a spoiler text visible only in other users' timelines (Zignani et al. 2019).

Mastodon distinguished between two timelines: a federated timeline and a local timeline. The local timeline shows posts by every user on the instance, whereas the federated timeline includes users from other instances if the user's home base is connected to it. Therefore, users can follow others from different instances (Farokhmanesh 2017), creating an inter-domain federated model similar to email (Raman et al. 2019). In a federated network, the entire network consists of decentralised nodes, each of which has local autonomy and can decide over its resources. Every node exists on its own server. This decreases the risk of network failure. The nodes can be connected as decided by the local administrators, who usually also own the servers on which the nodes are saved. These servers are called instances.

Instances can connect while staying independent, each having its own rules and purposes (Zulli et al. 2020). Users can either join existing instances or find an instance themselves by installing the necessary software. Instances can accept the registration of new users, making them open, or they can be closed, meaning new users can register only if they are invited by an administrator. Instances can use tags to show which topics and activities they allow. The majority of instances use the tags tech, games, and art. Furthermore, instances can specify which activities are allowed or prohibited on their instance. The most commonly prohibited activities are spam, pornography, and nudity (Raman et al. 2019). Mastodon contains a built-in function for users to mark posts as inappropriate or sensitive (Zignani et al. 2019).

Each instance has a different policy in dealing with content that is available to users upon registration. Mastodon.social, the biggest European instance, bans content illegal in France and Germany, for example, Nazi symbolism and Holocaust denial (Zignani et al. 2019). Mastodon.social specifies rules that might lead to deletion or suspension of users or toots if they, for instance, include illegal content, discriminate and harass, mob, or include violent or nationalist content. Moreover, they include best practices to be followed by users, such as the use of content warnings for "not safe for work" (NSFW) content, providing credit for creative work posted, or the avoidance of comments on others' appearances (mastodon, n.d.). These rules need to be accepted by users before registering for an instance. This increases transparency and trust in the administrator. Moreover, it is easier to justify decisions if the rules are known to users.

However, deletion is the last resort on mastodon.social. Most of the time, the administrators try to have a conversation with the user, applying a two- or three-strike approach before banning users or deleting content. If users do not comply with the rules and the administrators have the impression the situation will not get better, they might ask them to leave the instance. For a user from an external instance, the reported status and sometimes the profile and history of a remote user are checked to decide if a user is silenced, suspended or if no actions are taken. Generally, users from external instances are suspended because they are spammers, trolls, or they advocate for a right-wing ideology. Entire instances might also be blocked if they are led by right-wing or troll administrators. Suspending an instance leads to deleting the locally cached data, and a reset of all follows to that instance. The list of blocked instances is publicly available (List 2021). Overall, transparency is crucial for content moderation on mastodon.social. If moderation is not transparent and balanced, administrators might lose the users' trust. In that regard, a feeling for content that might hurt but not be harmful and for content that is not permissible, even as it might not hurt anyone, is necessary.

IV. Der Standard

STANDARD-Community is one of the largest German-language platforms for online debate. The community is managed by the Austrian newspaper Der STANDARD and works closely with its online community to manage and moderate their online forum. Key community members have met STANDARD-Community moderation team in person, and there are regular exchanges between the community and its members.

The content moderation on STANDARD-Community attempts to take a pre-emptive approach to content governance. In this approach, the moderation team is not primarily on the platform to delete content, but rather to identify challenging debates that are likely to go in direction that would need to be moderated and steer them clear of needing to be moderated. The primary focus of moderators and the moderation team is thus on promoting healthy debates rather than deleting problematic ones.

This approach to content governance also involves close collaboration with academics working in this area. Together with the Austrian Research Institute for Artificial Intelligence (OFAI), STANDARD-Community have developed an automated system to pre-filter problematic content, which pre-moderates content until it can be checked by a team of professional moderators. This approach has met some resistance within STANDARD-Community, as the pre-moderation of content is not always accurate and raises considerable challenges. STANDARD-Community also has an ongoing project with OFAI to increase the participation of women in STANDARD-Community.

Together with the Sustainable Computing Lab at the Vienna University of Economics and Business, STANDARD-Community developed a set of pre-emptive moderation techniques to prevent problematic content being posted in the first place. This involved a series of field experiments with A/B tested forum design changes. Whether users' posts in the A/B groups were deleted for breaking forum rules was used as a measure of the effectiveness of these forum design changes. These forum design changes culminated in a set of two videos that were presented to users, using social theory to influence user behaviour (Wagner and Kubina 2021). All of these forum design changes were openly and transparently communicated to forum users, so that they knew what was happening at every step of the way.

V. Slashdot

Slashdot is a social news website featuring stories about various topics, such as science, technology, and politics. Users can submit stories by using a submissions form. Subsequently, editors check and change the grammar, spelling, or formatting of the submission. Moreover, they review links and possibly remove or substitute them if they are not relevant or broken. According to the frequently asked questions (FAQ), the editors behind Slashdot choose the most relevant, interesting, and timely submissions for publication. Once published, other users can comment on the submitted stories (slashdot, n.d.).

Content governance on slashdot centres around strong user involvement in ranking and rating content, making it more or less visible according to a pre-standardised system. Furthermore, moderation decisions by users are judged by other users ensuring accountability. Slashdot pursues a strong policy of not deleting content but relying on making harmful content less visible.

On Slashdot, users are allocated so-called moderation points. They can use these points to moderate comments by other users by selecting an appropriate option from a drop-down list. The list includes the following options: Normal, Offtopic, Flamebait, Troll, Redundant, Insightful, Interesting, Informative, Funny, Overrated, and Underrated (slashdot, n.d.). Moderators are given five moderation points to be used within three days (Lampe and Resnick 2004). Users can spend their moderation points to promote or demote comments.

Users who are among the most active in the system are not chosen to moderate. Moreover, users cannot moderate the topics they comment on a lot. Such system is in place so people with extreme opinions, who tend to be more engaged, cannot be disruptive or moderate according to their own agenda. Moderators are also subject to moderation by the meta-moderation function. After the first moderation, other users ensure that the moderators have done a good job. Meta-moderators are users whose accounts are among the oldest, which are about 92.5% of Slashdot users. Administrators have unlimited moderation points (slashdot, n.d.).

Generally, no content is deleted from Slashdot. Administrators can ban some internet protocol (IP) addresses if they see abuse or spam. Moreover, users are banned if they are moderated down several times in a short time period (slashdot, n.d.). Users who get banned might make a new profile; thus, deleting content or banning users might incentivised them to act out, ultimately making the problem worse. As users are banned or downvoted, the instigators of trolling or other harmful behaviour have less influence; their followers still do, possibly causing harm.

Users collect karma, which represents how their comments have been moderated. Karma is improved if a user posts comments that are rated positively by moderators. If a post is moderated up, a user's karma goes up as well. If a user posts spam, their post is moderated down, and their karma decreases as well. Karma is structured on a scale that includes terrible, bad, neutral, positive, good, and excellent. Furthermore, an accepted story submission increases karma as well as good meta-moderation (slashdot, n.d.). Comments of users with high karma start at a higher score. Only users with high karma can be eligible for moderation (Lampe and Resnick 2004).

Users can set a threshold for the comments displayed to them. Comments are scored from -1 to 5, and users choose the threshold within that range. The higher the threshold for comments, the fewer comments a user can see (slashdot, n.d.). In its default view, the Slashdot forum is in a threaded structure with a threshold of +1. The direct responses to a story are shown in their entirety if they have a rating of +1 or higher. Responses are displayed in chronological order and indented. Further down the thread, only responses with a score of at least 4 are shown in their entirety, whereas comments rated 1–4 are partly shown, and comments below the threshold are omitted (Lampe, Johnston and Resnick 2007).

Lampe and Resnick (2004) analysed usage logs of comments, moderation, and meta-moderation during a two-month period in spring 2003. Their analysis of more than a million comments, moderations, and meta-moderations revealed that participation in the moderation system is quite high. About 28% of comments were rated at least once during the study, and 79% of moderations were positive. According to the study, disagreement between moderators and meta-moderators was scarce, and 15% of moderated comments received positive and negative feedback. A total of 92% of meta-moderations agreed with the decisions of the moderators. A conversation only needs a little more than an hour to reach its half-life, and moderation should support users to allocate their attention. Therefore, moderation needs to happen relatively fast. The study found that comments received moderation in a median time of 83 minutes. However, pushing a comment to a +4 or -1 score took almost three hours. Furthermore, the study found that comments with a higher starting score received moderation sooner. The authors suggested that comments should be moderated quickly and accurately.

Moreover, each moderator should only have a limited impact on a comment, and to ensure moderators' participation, their workload should be minimal (Lampe and Resnick 2004). On Slashdot, two to five moderators are required to rate a comment positively for it to reach the highest score. That limits the influence of individual moderators. It takes, however, almost three hours, which is too long to be effective. The moderators choose which comments to rate. Although this might reduce their effort, it also leads to biases, as comments with higher initial scores are moderated faster, and others, which are made later in the thread, receive lower scores than deserved (Lampe and Resnick 2004).

VI. Advantages and disadvantages of different approaches

Each of the outlined approaches to content governance provides its own set of advantages and disadvantages, depending on the purpose of the platform and the goals of the content moderation. Overall, a combination of different approaches seems to be encouraged by participants, depending on the scale of the platform, the use of automated systems, and the human resources available. The table below summarises in detail the advantages and disadvantages of content moderation approaches as used on the investigated platforms diaspora* (d*), mastodon (m), Wikipedia (W), slashdot (/.) and Der Standard (DS).

TABLE 1: ADVANTAGES AND DISADVANTAGES OF DIFFERENT CONTENT GOVERNANCE APPROACHES

APPROACH TO CONTENT GOVERNANCE	ADVANTAGES	DISADVANTAGES	USED ON
Deletion of content	• Harmful content is no longer accessible • Illegal content is removed • No copycats • Not giving a platform to undesired content	• Might inspire trolls to post more • Might delete content that shows human rights violations • Might harm freedom of expression	W, m, d*, DS
Ban/ suspend user accounts after repeated offense (Deplatforming)	• Might be temporary and reversible if mistakes are made • Possibility to reform user behaviour/ enforce compliance • Spammers or vandals are removed	• Trolls/Bullies might make a new account • Users might set up their own pod/ space • Even though the perpetrator is removed, followers might continue	/., W, m
Conversation/ Warning Users	• Possible to clear misunderstandings • Use of open discussion pages about issues • Could enforce civility	Need a lot of administrators resources	m, d*, DS, W
Moderation by users (blocking/ flagging/ points)	• Less resources for administrators • Reduces problems of scale • Might strengthen the community	• Might be too slow • Careless users or bullies might be unfair • Could be gamed	/., W, m, d*, DS
Inclusion of Meta-Moderation	• Accountability of moderators can be secured • Strengthens the community • Admins' rights can be removed if they lose community trust	• Biases by meta-moderators • Might take too long	/., W
Hiding Content	• Harmful content is not accessible anymore • Flexible and reversible	Poster might realise and might repeat their harmful posts multiple times	d*, W
Downvoting/Upvoting Content	• Harmful content is less visible • The community decides on the content that is included	Making harmful content more visible by organised trolls/bullies	/.

Automated filters with human review	• Reduces false positives • Easier to find potentially harmful content • Reduction of workload • Temporarily blocking suspicious content	• Psychological burden of seeing/ reading harmful content • Harm to freedom of expression while content is waiting for moderation	m, W, DS
Automated filters without human review	• Easy way to get rid of spam • Human moderators do not need to see harmful content	• False positives • Computers don’t understand context/subtexts • These types of algorithms are typically biased	W
Blocking Instances/ nodes/ pages	Get rid of instances with troll/spam admins	New instances or pages might be created	m, W, d*
Use of content warnings (Spoiler, NSFW)	• Protects users from content they do not want to see • Administrators need to take down/ hide/review less content	Trolls or people that deliberately post harmful content would not use them	m, W
Making rules known in advance	• Decreases mistakes and misunderstandings • Easier to justify administrators’ decisions to block/delete	Trolls might just ignore them	m, W, DS
Edits/ Content that is added needs to be approved first/ pre-moderation	Harmful content can be filtered out before it is posted	• Significant resources by administrators needed • Biases by administrators • Users might not know why their content is not posted	W, DS
Karma points/ Reputation of users	• Incentive for good behaviour/contributions • Only users with high karma can be selected for moderation	Users with a high score are more visible, making it more difficult for new users, who don’t have a high score yet	/., W

Concrete alternatives based on empirical data

We have investigated alternative content governance models to assess their effects on online communities. The following sections provide an overview of a selection of four alternative content governance approaches that have shown promising outcomes based on concrete empirical evidence. The approaches covered include an accountability approach to increase rule compliance, real-name approaches, ranking and rating as well as forum design changes to increase rule compliance.

I. Accountability approach to increase rule compliance

Most online platforms do not divulge how they reach content moderation decisions and resort to deleting content without providing the user with an explanation. The lack of transparency in how online platforms make these decisions leads to a sparsity of good data (Suzor, Van Geelen, and Myers West 2018; Jhaver et al. 2019b), which complicates research efforts (Jhaver et al. 2019b, 4). Prior research indicates that the opacity of the content moderation processes induces users to develop "folk theories" about how and why their content was removed (Eslami et al. 2015; Jhaver et al. 2019a). Thus, the complexity of the content moderation infrastructure and processes makes content regulation incomprehensible for end users and risks diminishing its legitimacy (Jhaver et al. 2019b).

There is a growing field of research on strategies beyond merely sanctioning undesirable content or bad actors to improve an online community's health (Jhaver, Vora, and Bruckman 2017; Mathew et al. 2019). One of these alternative strategies consists of increasing the transparency of the moderation process by providing explanations to end users about why their content was removed. The centrality of explanations for system transparency has previously been demonstrated in areas as diverse as e-commerce (Pu and Chen 2006; Wang and Benbasat 2007), medical recommendation systems (Armengol, Palaudaries, and Plaza 2001), and data exploration systems (Carenini and Moore 1998).

In an award-winning paper applying topic modelling techniques to a corpus of 22,000 removal explanations on Reddit, Jhaver et al. (2019b) explored how transparency in moderation at different levels affects subsequent user behaviours using a sample of including future post submissions and future post removals. Moderators on Reddit can provide removal explanations to users either by commenting on a removed post, sending a private message to the offending member, or highlighting removed posts with a short tag (so-called 'flairing').

Jhaver et al. (2019b) contend that explanations help Reddit users learn the norms of the Reddit community in the three ways identified by Kiesler et al. (2012), that is, through direct feedback, visible community guidelines, and the observation of other people's behaviour and consequences thereof. First, the authors found that explanations help post submitters learn the norms of the community by providing direct feedback from the moderator team on how their submissions did not align with the norms of the community (Jhaver et al. 2019b, 20). Second, given that explanation messages usually indicate the community norm the submission violated, often coupled with a link to the wiki page for the subreddit rules, they help users better understand the explicit social norms of the community and increase the likelihood of rule compliance. Lastly, according to the authors, as explanation messages are posted publicly, they are visible to many users, informing bystanders through observation of why certain types of posts are unacceptable in the online community.

The analysis of the data reveals that providing explanations for post removal decreases the likelihood of future post removal of moderated users, whereas simple deletion without an explanation actually increases the odds that a user will experience post removal in the future. The authors of the study calculated that future post removals could be reduced by more than 20% when all removals on Reddit were to be paired with an explanation, implying a considerably lower workload for moderators (Jhaver et al. 2019b, 21). Explanation comments are more elaborate, going beyond the context of the current removal and containing additional information on a potential appeal process if the user considers the decision a mistake. This might explain why explanatory comments are more effective at preventing future removals than tagging posts with short flairs (Jhaver et al. 2019b, 22). Flairs are, however, more commonly used, probably because explanation comments are more time-consuming (87% versus 11%).

Jhaver et al. (2019b) concluded that moderators should be encouraged to take the time to provide explanation comments instead of flairs. They also found that providing explanations for content removal could also be integrated into automated tools, as the explanations provided by bots and other automated tools were found to be effective in preventing future post removals (Jhaver et al. 2019b, 22). They cautioned, however, that automated tools are prone to making mistakes, which could jeopardize users' trust in the moderation process. Accordingly, automated tools need to be designed and deployed carefully, and human moderators should continue to provide explanations in cases where reasons for removal are unclear (Jhaver et al. 2019b, 23).

II. Durable pseudonyms

People have been found to behave differently when they believe their behaviour is visible to others and can be attributed to them. For example, people more readily admit socially unsanctioned activities or controversial points of view in anonymous surveys compared to in-person interviews (Richman et al. 1999) or give more money to charities in public than in an anonymous setting (Alpizar, Carlsson, and Johansson-Stenman 2008). Although these studies indicate that identifiable individuals may be more inclined to follow social norms than anonymous individuals, more recent research suggests that previous studies might have defined the concept of anonymity too broadly, ignoring other factors, such as the importance of eye contact or lack thereof (Lapidot-Lefler and Barak 2012).

While much of the public debate draws a strict dichotomy between anonymity and real-name identity, the reality is slightly more complex. In fact, there is a broad spectrum of identity disclosure on online platforms, from the purest form of anonymity in the form of contributions that cannot be traced back to any particular commenter and pseudonyms allowing a persistent identity in a forum to online platforms that require the use of one's legal name and various other configurations in-between. However, requiring a legal name leads to considerable problems from a human rights perspective. This is because requiring a legal name tends to exclude marginalised groups that are frequently less able to communicate online using their real name and are more likely to face repercussions for expressing themselves without breaking any rules.

Durable pseudonyms can have considerable benefits for the quality of debate without the human rights challenges that real name policies bring with them. By contrast, *cheap pseudonyms* refer to the ease with which an online identity can be acquired and changed. Cheap pseudonyms allow users to violate community norms without paying reputational consequences, such as bans, because they can easily create a new account name and continue to perpetrate their offending behaviour. To increase incentives for users to maintain a pseudonym when they are sanctioned, the continued use of the current identity, and thus accepting the sanction, must be more attractive than creating a new pseudonym (Friedman and Resnick 2001).

Kiesler et al. suggest three strategies to maintain effective sanctions, given the possibility of cheap pseudonyms. First, it is important to increase the appeal to keeping a long-term pseudonym, for example, by requiring a certain seniority for certain capabilities, or by giving more weight to contributions of long-term members (Kiesler et al. 2012, 159). Wikipedia, for example, restricts voting rights in the Wikimedia Stewards Elections to editors with an edit count of at least 600 overall and 50 in the last six months (Wikipedia 2021f).

Second, maintaining long-term pseudonyms can confer prestige. For example, in communities where user IDs are assigned sequentially, early accounts with low ID numbers are seen as valuable. For example, Slashdot included a low user ID as one of the items in a charity auction (Kiesler et al. 2012, 159). With such a mechanism in place, a member with a low user ID might be more inclined to accept sanctions to avoid having to create a new account with a high ID. Importantly, long-term pseudonyms still allow a considerable degree of anonymity when using them on a platform.

For example, a long-term pseudonym used on a platform like Wikipedia, on which a user has built up a positive reputation and become an editor, may take many years to maintain. This reputation can be built up without the need for any personally identifying information to Wikipedia.[4] The user of such a long-term durable Wikipedia account cannot easily recreate it without years of work. Thus, the consequences of a decision to sanction this user or even ban them permanently are highly consequential for this user. As such, durable pseudonyms do not have many of the downsides associated

[4] See https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_anonymous for further details.

with real-name policies, which require the disclosure of personally identifying information. Thus, durable pseudonyms can be considered a highly preferable human rights-friendly alternative to real-name policies.

Lastly, a long-term pseudonym can be designed to yield financial benefits. For example, the high reputation of an eBay account was shown to generate 8% more selling revenue compared to a new account (Resnick et al. 2006). However, when such durable pseudonyms have financial benefits attached, their effects on individual human beings' livelihoods should also be taken into account when restricting access to them.

III. Ranking and rating

Another strategy to build trust, promote good behaviour and encourage contributions relevant to the community is to resort to reputation systems. Reputation systems are ubiquitous in e-commerce websites, such as eBay and Amazon.com. In these settings, where a large number of strangers interact for the first time, a reputation system allows a buyer to gauge the trustworthiness of a seller based on the feedback given by previous customers.

Reputation systems summarise the history of someone's online behaviour into one or several reputation values, which are displayed in an actor's profile, enabling users to establish trust based on the experiences made by others (Resnick et al. 2000). Reputation can be user-generated (or explicit), provided by users who have interacted with the person being rated, or it can be system-generated (or implicit), computing a score from a user's activity on a site.

In e-commerce settings, reputation systems therefore facilitate decision-making, leveraging the collective work of other users to reduce the costs of evaluating which user is trustworthy (Lampe 2011). In addition to e-commerce websites, variations of this strategy have been adopted by a number of online communities. While each site uses a slightly different reputation system, they generally track the behaviour of members by giving users "karma" points for their posts and other activities, as well as the ability to upvote (and, usually, also downvote) other's contributions. When a post is upvoted or downvoted by fellow members of a community, the poster receives or loses points.

Systems vary according to whether a site allows *upvotes* and *downvotes* or just *upvotes* and the weight attributed to each. Furthermore, some sites display the reputation score next to the user's name (thus serving as an indicator to the community), whereas others use it primarily to tell the site's algorithm which posts are interesting enough to go on the front page.

Research has shown that a particular community's reputation and rating system shape how users interact with each other (Lampe and Resnick 2004; Lampe 2006). On Reddit, for example, users have unlimited upvotes and downvotes per day, with no consequences of using them. This lack of reputational cost to downvotes encourages more casual and oftentimes incendiary comments. A number of subsequent communities have sought to avoid this dilemma entirely by severely limiting downvotes (Hacker News requires a karma threshold of 100), whereas communities like Stack Overflow have designed the weight of both upvotings and downvotings to encourage "*responsible* downvoting" (Atwood 2009). Slashdot randomly gives users a certain number of points they can use for upvoting or downvoting. Therefore, users use their allocated moderation points with more consideration. One of the main downsides of rating and ranking systems is their vulnerability to gaming, both by bots and trolls. It is thus particularly important to take this into account when implementing a system of this kind and ensure that measures are in place to systematically guard against these vulnerabilities.

In addition to indicating which contributions are worth reading, reputation and rating systems also serve as socializing functions by providing explicit feedback from other users (Lampe 2011, 78). This feedback enables a user to learn about a community's norms and, at the same time, enforces said norms. For example, a survey of new members of the community, Slashdot, showed that new users studied the ratings of contributions before posting their own. The same study found that the longer new users wait before posting a first comment, the more positive their first contributions would be rated. This indicates that the feedback provided by the rating system allowed users to see what type of content was valued on the site before posting themselves (Lampe and Johnston 2005).

IV. Forum design changes

The design of an online forum can have a considerable impact on how debates within the forum take place. Interviews conducted during this study suggest that there is considerable scope for reshaping online platforms to ensure greater compliance with forum rules. However, data on these kinds of changes to online platforms are rarely published and, thus difficult to assess from an external perspective. One of the few communities that was able to provide this data is STANDARD-Community, providing us with reliable data on the effects of forum design changes on rule compliance in online forums.

The recently published blog post with preliminary results of a video experiment by STANDARD-Community is of particular interest in this regard (Wagner and Kubina 2021). In this field experiment conducted as part of an A/B-test in January 2020, users of the STANDARD forum were split into three groups at random: Group A, which was shown Video A; Group B, which was shown Video B; and Group C, which was a control group. After seeing the video, the users completed an online questionnaire to better understand why they follow rules in the first place.

Video A primarily focused on the negative consequences of users' actions, reminding users that there could be legal consequences if they broke forum rules. Video B focused on defining appropriate behaviour in online forums, within which the whole group suffers if a small group of users did not follow the rules. Group C was not shown a video as part of the control group, but was still asked to fill out a questionnaire.

The results of this field experiment showed that there was no difference between the control group and the users who saw video A. By contrast, users who saw Video B produced approximately 19% less content that needed to be deleted by moderators than the control group. In public policy terms, this suggests that threatening users with ever more draconian legal sanctions is unlikely to be successful in reducing rule-breaking behaviour in online forums. By contrast, using changes in forum design to promote different forms of appropriate behaviour within specific communities may be particularly effective in getting users to change their behaviour. Although deletion remains necessary, it should be used as a last resort rather than as the primary mechanism of content governance.

Policy recommendations

Policymaking in this area is extraordinarily difficult. Democratic governments have constitutional limitations in many areas and should often regulate them to safeguard freedom of expression. At the same time, there is indisputable harm caused by the many types of content currently available online. When regulation is stuck between a rock and a hard place, all policy measures need to be carefully targeted, and their impacts measured to ensure that they are effective in achieving their stated goals. States need to ensure that their policy measures respect, protect, and enable all human rights. During the course of the interviews conducted for this report, we compiled a wide variety of ideas and suggestions on how to develop better policy. Many of these recommendations have been integrated into the following text, together with the recommendations of the authors on how to proceed.

1. **Deleting content is not a solution**; it is simply a 'Band-Aid' for an already existing problem. When online communities produce large amounts of content that need to be deleted, the production of this content itself is the problem.

2. **The quality of content moderation processes is key** to achieving human rights-centred content moderation models that empower users. Regulatory proposals that push platforms to automate content moderation by requiring ever shorter timeframes for their response (24 hours, 2 hours, etc.) will continue to miss its mark. Instead, push platforms to innovate in ways that change leads to better content moderation and better content on online platforms.

3. **Automating content moderation is not a solution.** Rather, automation can support certain limited areas of content moderation and content governance. Unduly short time frames for content removal create unhelpful pressure for platforms to use highly problematic automated tools. One interviewee mentioned a "cottage industry of junk technology" being pushed on smaller platforms without the resources of large tech giants. It is thus very dangerous for public regulators to mandate automation or very fast response times, as the knock-off effect will promote junk tech rather than human rights.

4. **Meaningful human involvement in content moderation decisions is key** to effective content moderation. Content moderation staff with training, time, and capacity are needed to ensure effective content moderation in online platforms. These staff do not necessarily need to be employees; notably, our research suggests that involving trustworthy community members in content moderation practices strengthens the legitimacy of content moderation decisions.

5. **Democratise platforms' terms of services.** As the terms of service include many provisions that go beyond what is legally required for platforms to moderate, it is important that these terms of service are jointly developed, with the community being moderated. Our research suggests that jointly developed and clearly communicated ToS, regardless of their content, are more likely to be adhered to and seen as legitimate by community members than unilaterally ToS imposed by platforms. Participation in ToS development is a crucial element in what makes community-led and community-driven platforms successful.

6. **Platforms need to clearly communicate to those affected** by their decisions what kind of moderation techniques the platform is using, what tools are being used, and provide a redress mechanism in case mistakes are made. Different content moderation tools are being applied at different levels of content moderation without sufficient transparency.

7. **Community-led and community-driven content moderation seems to be highly effective,** based on the research and data provided in this study. This is due to the fact that community moderation offers "alternative models of top-down moderation," in contrast to commercial content moderation on platforms that mainly rely on assessing and deleting a single piece of content.

8. **Platforms need to build on best practices for content moderation and create mechanisms for sharing them.** During our interviews, numerous such practices were emphasised by content moderators as examples of a good practice. For instance, a 'scale of power' assigned to individual users based on their interactions and credibility proved to be an effective tool that promotes the sense of community and positively shapes interactions among others. Another mentioned example of good practice is the increased visibility of high-quality comments.

9. **Systematically audit content moderation and content governance policies.** At present, there is insufficient knowledge available on the effectiveness, false positives/false negatives, harms, and human rights compliance of different content moderation and governance policies. Systematically auditing these policies and sharing the results could enable an ecosystem that improves policies rather than the current race to the bottom (Wagner et al. 2021). Depending on the purpose of the audit and the system audited, different approaches can be appropriate, including code reviews or the examination of training data. It is essential that auditing involves all relevant stakeholders with required expertise to analyse a) what policies platforms have in respect to different categories of content; b) what interventions and processes that the platforms follow, including both technical and personnel implementation; and c) the outcomes of these processes and how they impact users' fundamental rights and freedoms.

10. **Support community-oriented platforms in building digital public spaces.** In particular, non-profit and community-oriented platforms, as well as open-protocol and networks for decentralised communication, deserve financial support in developing their platform moderation efforts. This financial support should be contingent on their willingness to continue providing valuable digital public spaces and making the results of their efforts publicly availability. Healthy, critical and inspiring debates are in the interest of societies as a whole.

11. **Meaningful transparency measures** have to be an integral part of any content moderation system. Transparency builds trust between users and content moderators. Moderation practices significantly vary among all but the largest online platforms. Platforms should publish aggregated data on how their content moderation teams are being built up. Such data should include age, nationality, race, gender, and linguistic skills of content moderators as well as whether and how often they receive human rights training. Platforms should also explicitly separate legally-required content moderation from content moderation that goes beyond the legal requirements (i.e., ToS) in their transparency reports. Particularly, information about guidelines used by content moderators and what processes exist to support moderators in making consistent decisions should be disclosed to users.

12. **Contribute to diversity in spaces for public debate.** The largest online platforms have relatively similar moderation practices, leading to the danger of creating content cartels (Douek 2021) that reinforce a global default of permissible speech (Wagner 2016). To counter this trend, it is essential for the future of human rights-centric platform governance in the EU that lawmakers support and enable decentralised and community-led platforms that will protect and empower individuals online.

13. **Effective and easily accessible accountability measures and redress mechanisms** need to be in place. This includes a notification that should take place before any action is taken against flagged or notified content and should contain adequate explanation of what rule was breached, how, and what next steps will be taken with regard to the piece of content, introducing the safeguard of procedural fairness, and redress mechanisms. Users should be provided with meaningful explanations of how breaches of rules are being identified and enforced.

14. **Build a sustainable research ecosystem on content moderation and content governance in Europe.** At present, there is insufficient empirical research on content moderation and content governance in Europe. What little research exists in this area takes place in United States-based private companies and is rarely publicly available.

15. **Protect freedom of expression for all.** Online platforms provide public spaces for broad online debate. Crucially, these spaces need to enable free expression so that all members of society can safely engage in them, regardless of skin colour, gender or sexual preferences. A public space is not a public space if numerous parts of society have reasonable grounds to fear engaging in it.

16. **Improve support, protection, and training of content moderation staff.** Currently, there are no common standards on minimum levels of support, protection, and training for human beings in online content moderation. Even for the worst types of content, we are familiar with cases (outside the scope of this study) where insufficient support and protection are provided to those staff. Platforms should be required to have a basic standard of care for these staff, as well as for any volunteers engaged in content moderation. To this end, a clear category of volunteer content moderators with defined the rights and responsibilities should be created.

Final considerations

Online platforms do not mirror society. Rather, each platform creates its own logic of appropriate content governance. The result is many communities with different types of appropriate content in them. To blame failures in society for the failures of platforms in setting standards of appropriate content is, therefore, neither reasonable nor helpful in solving real human rights violations nor the harms created through online platforms.

A large body of research has focused exclusively on platform-driven moderation under centralised models of dominant platforms without considering alternative moderation practices. The outcomes of our study showed that alternative approaches to content governance that are the part of community-structured platforms allow for more nuanced and context-sensitive moderation.

In this study, we have attempted to **reimagine content moderation** in online platforms. The current situation in which an invisible moderator located in an unknown place makes unaccountable decisions about the boundaries of public spaces is not sustainable. There is a need for much better content moderation and content governance practices that do not rely on deleting content as a solution but perceive it as an uncomfortable necessity, which demonstrates a wider problem on the platform removing the content without any accountability.

Community-led and community-driven platforms demonstrate every single day that a different way of moderating content is possible. These platforms tend to involve their communities in moderation decisions as well as in developing the ToS based on which these decisions are made. This community-embedding promotes greater scrutiny, legitimacy, and effectiveness of community rules and decisions, while ensuring that less content needs to be moderated in the first place.

Thus, taking a **holistic perspective on content moderation is key**. For content moderators at scale, it is easier and faster to push away the content that needs to be moderated and, consequently, delete, block or disabled its presence on the platforms. It is far more difficult to acknowledge that the content produced at the edges of the permissibility of an online community reflects on the nature of the online community itself. Online platforms cannot look away from this content, delete it, or pretend that it does not exist. Instead, they should acknowledge that this content is also part of their community and reduce the likelihood that it will be produced in the first place.

This more holistic approach to online content moderation can contribute to **humanising online platforms** (Ruckenstein and Turunen 2020) and safeguarding fundamental rights. Online platforms are not just used for humour, cute babies, and cat pictures. They are key public spaces that influence public debate and human behaviour. This makes it so important that democratic governments take on the challenge of enabling the many platforms that will be needed to promote diverse public debate in democratic societies.

References

Ahlert, Christian, Chris Marsden, and Chester Yung. 2004. 'How "Liberty" Disappeared from Cyberspace: The Mystery Shopper Tests Internet Content Self-Regulation'. Programme in Comparative Media Law & Policy at the Oxford Centre for Socio-Legal Studies Research Paper. https://www.academia.edu/686683/How_Liberty_Disappeared_from_Cyberspace_The_Mystery_Shopper_Tests_Internet_Content_Self_Regulation

Alpizar, Francisco, Fredrik Carlsson, and Olof Johansson-Stenman. 2008. 'Anonymity, Reciprocity, and Conformity: Evidence from Voluntary Contributions to a National Park in Costa Rica'. Journal of Public Economics 92 (5–6): 1047–60.

Armengol, Eva, Albert Palaudaries, and Enric Plaza. 2001. 'Individual Prognosis of Diabetes Long-Term Risks: A CBR Approach'. Methods of Information in Medicine-Methodik Der Information in Der Medizin 40 (1): 46–51.

Atwood, Jeff. 2009. 'The Value of Downvoting, or, How Hacker News Gets It Wrong'. Stack Overflow Blog. 9 March 2009. https://stackoverflow.blog/2009/03/09/the-value-of-downvoting-or-how-hacker-news-gets-it-wrong/.

Berríos, Reinaldo, and Nydia Lucca. 2006. 'Qualitative Methodology in Counseling Research: Recent Contributions and Challenges for a New Century'. *Journal of Counseling & Development* 84 (2): 174–86. doi: https://doi.org/10.1002/j.1556-6678.2006.tb00393.x.

Blackwell, Lindsay, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. 'Classification and Its Consequences for Online Harassment: Design Insights from HeartMob'. In *Proceedings of the ACM on Human-Computer Interaction* 1 (CSCW): 24:1-24:19. doi: 10.1145/3134659.

Bordia, Prashant. 1997. 'Face-to-Face Versus Computer-Mediated Communication: A Synthesis of the Experimental Literature'. *Journal of Business Communication* 34 (1): 99–118. https://doi.org/10.1177/002194369703400106.

Braun, Virginia, Victoria Clarke, and Debra Gray, eds. 2017. *Collecting Qualitative Data: A Practical Guide to Textual, Media and Virtual Techniques*. Cambridge: Cambridge University Press.

Caplan, Robyn. 2018. *Content or Context Moderation: Artisanal, Community-Reliant and Industrial Approaches*. New York, NY, USA: Data & Society Institute.

Carenini, Giuseppe, and Johanna Moore. 1998. 'Multimedia Explanations in IDEA Decision Support System'. In Working Notes of the AAAI Spring Symposium on Interactive and Mixed-Initiative Decision Theoretic Systems, 16–22.

Dara, Rishabh. 2011. 'Intermediary Liability in India: Chilling Effects on Free Expression on the Internet'. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2038214.

Davies, Todd, and Reid Chandler. 2011. 'Online Deliberation Design: Choices, Criteria, and Evidence'. In *Democracy in Motion: Evaluating the Practice and Impact of Deliberative Civic Engagement*, edited by Tina Nabatchi, John Gastil, G. Michael Weiksner, and Matt Leighninger, 103–31. Oxford: Oxford University Press.

De Laat, Paul B. 2014. 'The Use of Software Tools and Autonomous Bots Against Vandalism: Eroding Wikipedia's Moral Order?' *Ethics Inf Technol* 17, 175–188, doi: 10.1007/s10676-015-936its.6-9.

De Streel, Alexandre, Elise Defreyne, Hervé Jacquemin, Michèle Ledger, and Alejandra Michel. 2020. 'Online Platforms' Moderation of Illegal Content Online', 102.

diaspora*, n.d.a. Retrieved January 28, 2021 from https://diasporafoundation.org/.

diaspora*, n.d.b. Retrieved January 28, 2021 from https://discourse.diasporafoundation.org/.

Douek, Evelyn. 2021. 'The Rise of Content Cartels'. New York, NY: Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/the-rise-of-content-cartels.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. 'I Always Assumed That I Wasn't Really That Close to [Her]: Reasoning about Invisible Algorithms in News Feeds'. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–62.

Farokhmanesh, Megan. 2017. A beginner's guide to Mastodon, the hot new open-source Twitter clone. The Verge April 7. Retrieved on February 17, 2021 from https://www.theverge.com/2017/4/7/15183128/mastodon-open-source-twitter-clone-how-to-use

Friedman, Eric J., and Paul Resnick. 2001. 'The Social Cost of Cheap Pseudonyms'. *Journal of Economics & Management Strategy* 10 (2): 173–99. https://doi.org/10.1162/105864001300122476.

Gerdes, Daniel A., and James Conn. 2001. 'A User-Friendly Look at Qualitative Research Methods'. *The Physical Educator* 58(4).

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Github. Chaossocial/about. List of instances blocked by chaos.social. Retrieved on February 18, 2021 from https://github.com/chaossocial/about/blob/master/blocked_instances.md

Goldman, Eric. 2021. *Content Moderation Remedies*. Working Paper. Santa Clara University School of Law. U.S.

Hanna, Paul. 2012. 'Using Internet Technologies (Such as Skype) as a Research Medium: A Research Note'. *Qualitative Research* 12(2):239–42. doi: 10.1177/1468794111426607.

Hille, Sanne, and Piet Bakker. 2014. 'Engaging the Social News User: Comments on News Sites and Facebook'. *Journalism Practice* 8 (5): 563–72. https://doi.org/10.1080/17512786.2014.899758.

Jhaver, Shagun, Pranil Vora, and Amy Bruckman. 2017. 'Designing for Civil Conversations: Lessons Learned from ChangeMyView'. Georgia Institute of Technology.

Jhaver, Shagun, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019a. 'Did You Suspect the Post Would Be Removed?: User Reactions to Content Removals on Reddit'. *Proceedings of the ACM on Human-Computer Interaction*, 31 (CSCW): 192:1-192:33. doi: 10.1145/3359294

Jhaver, Shagun, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019b. '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.

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert. 2019. 'Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit'. In *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–27. https://doi.org/10.1145/3359252.

Jowett, Adam. 2020. 'Carrying out Qualitative Research under Lockdown – Practical and Ethical Considerations'. *Impact of Social Sciences*. Retrieved 11 January 2021 (https://blogs.lse.ac.uk/impactofsocialsciences/2020/04/20/carrying-out-qualitative-research-under-lockdown-practical-and-ethical-considerations/).

Kiesler, Sara, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. 'Regulating Behavior in Online Communities'. pp. 125–78 in *Building Successful Online Communities: Evidence-Based Social Design*. Cambridge, MA: MIT Press.

Kruse, Jan. 2014. *Qualitative Interviewforschung: Ein Integrativer Ansatz*. Beltz Juventa; Auflage: 1.

Lampe, Cliff and Resnick, Paul. 2004. 'Slash(dot) and Burn: Distributed Moderation in Large Online Conversation Space'. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '04*). Association for Computing Machinery, New York, NY, USA, 543–550. doi: 10.1145/985692.985761.

Lampe, Cliff, and Erik Johnston. 2005. 'Follow the (Slash) Dot: Effects of Feedback on New Members in an Online Community'. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work - GROUP '05*, 11. Sanibel Island, Florida, USA: ACM Press. https://doi.org/10.1145/1099203.1099206.

Lampe, Cliff. 2006. 'Ratings Use in an Online Discussion System: The Slashdot Case'. Ann Arbor, MI: University of Michigan.

Lampe, Cliff, Johnston, Erik and Resnick, Paul. 2007. 'Follow the reader: Filtering Comments on slashdot'. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '07*). Association for Computing Machinery, New York, NY, USA, 1253–1262. doi: 10.1145/1240624.1240815.

Lampe, Cliff. 2011. 'The Role of Reputation Systems in Managing Online Communities'. In *The Reputation Society: How Online Opinions Are Reshaping the Offline World*, edited by Hassan Masum and Mark Tovey. The Information Society Series. Cambridge, Mass: MIT Press.

Lapidot-Lefler, Noam, and Azy Barak. 2012. 'Effects of Anonymity, Invisibility, and Lack of Eye-Contact on Toxic Online Disinhibition'. *Computers in Human Behavior* 28 (2): 434–43. https://doi.org/10.1016/j.chb.2011.10.014.

Leader Maynard, Jonathan, and Susan Benesch. 2016. 'Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention'. *Genocide Studies and Prevention: An International Journal* 9 (3): 8. doi: 10.5038/1911-9933.9.3.1317

Leah & Rix. 2020. Bericht: Sparschwein 2020. Retrieved on February 18, 2021 from https://blog.chaos.social/2020/01/26/sparschwein-bericht-2020.html.

Leshed, Gilly. 2009. 'Silencing the Clatter: Removing Anonymity from a Corporate Online Community'. In *Online Deliberation: Design, Research, and Practice*, edited by Todd Davies and Seeta Peña Gangadharan, 223–32. Stanford, CA: CSLI Publications.

Lessig, Lawrence. 1999. *Code: And Other Laws of Cyberspace*. New edition. New York: Basic Books.

Lo Iacono, Valeria, Paul Symonds, and David H. K. Brown. 2016. 'Skype as a Tool for Qualitative Research Interviews'. *Sociological Research Online* 21 (2): 103–17. doi: 10.5153/sro.3952.

Leyden, John. 2004. 'How to Kill a Website with One Email'. 14 October 2004. https://www.theregister.com/2004/10/14/isp_takedown_study/.

Mastodon. n.d. Rules. Retrieved on February, 18 2021 from https://chaos.social/about/more.

Mathew, Binny, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. 'Thou Shalt Not Hate: Countering Online Hate Speech'. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019),*13: 369–80.

Matias, J., Mou, Merry, Penney, Jonathon, Klein, Maximilian, and Wright, Lucas. 2020. 'Do Automated Legal Threats Reduce Freedom of Expression Online? Preliminary Results from a Natural Experiment'. https://doi.org/10.17605/OSF.IO/NC7E2.

McCracken, Grant. 1988. *The Long Interview*. 2455 Teller Road, Newbury Park California 91320 United States of America: SAGE Publications, Inc.

Merz, Manuel. 2019. Die Wikipedia-Community. Typologie der Autorinnen und Autoren der freien Online-Enzyklopädie. Wiesbaden: Springer.

Mills, Albert, Gabrielle Durepos, and Elden Wiebe, eds. 2010. 'Most Different Systems Design'. in *Encyclopedia of Case Study Research*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc.

Moses, Jonathon, and Torbjørn Knutsen. 2007. *Ways of Knowing: Competing Methodologies in Social and Political Research*. Macmillan Education UK.

Osman, Maddy. 2021. 'Wild and Interesting Facebook Statistics and Facts (2021)'. Kinsta. 3 January 2021. https://kinsta.com/blog/facebook-statistics/.

Penney, Jonathon. 2019. 'Privacy and Legal Automation: The DMCA as a Case Study'. *Stanford Technology Law Review* 22 (1): 412.

Pirkova, Eliska. 2016. 'Combating Hate Speech through Counter-Terrorism Measures: Can Words Harm?' In *Power Dynamics within Identity-Building: Revisiting Concepts and Paradigms*. Inter-Disciplinary Press.

Pallero, Javier, and Eliska Pirkova. 2020. '26 Recommendations on Content Governance: A Guide for Lawmakers, Regulators, and Company Policy Makers. Access Now.

Poduptime. 2021. Retrieved January 28, 2021 from https://diaspora.podupti.me/.

Pu, Pearl, and Li Chen. 2006. 'Trust Building with Explanation Interfaces'. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, 93–100.

Raman, Aravindh, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 'Challenges in the Decentralised Web: The Mastodon Case.' In Proceedings of the Internet Measurement Conference, 217–29. Amsterdam Netherlands: ACM, 2019. doi: 10.1145/3355369.3355572.

Resnick, Paul, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. 'Reputation Systems'. *Communications of the ACM* 43 (12): 45–48. https://doi.org/10.1145/355112.355122.

Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. 'The Value of Reputation on eBay: A Controlled Experiment'. *Experimental Economics* 9 (2): 79–101. https://doi.org/10.1007/s10683-006-4309-2.

Richman, Wendy L., Sara Kiesler, Suzanne Weisband, and Fritz Drasgow. 1999. 'A Meta-Analytic Study of Social Desirability Distortion in Computer-Administered Questionnaires, Traditional Questionnaires, and Interviews.' *Journal of Applied Psychology* 84 (5): 754–75. https://doi.org/10.1037/0021-9010.84.5.754.

Rings, Guido, and Sebastian Rasinger, eds. 2020. 'Theoretical Approaches'. pp. 83–202 in *The Cambridge Handbook of Intercultural Communication, Cambridge Handbooks in Language and Linguistics.* Cambridge: Cambridge University Press.

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press.

Roberts, Sarah T. 2017. 'Content Moderation'. In *Encyclopedia of Big Data*, edited by Laurie A. Schintler and Connie L. McNeely, 1–4. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_44-1.

Rogers, Richard. 2020. 'Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media'. *European Journal of Communication* 35 (3): 213–229. doi: 10.1177/0267323120922066.

Rowe, Ian. 2015. 'Civility 2.0: A Comparative Analysis of Incivility in Online Political Discussion'. *Information, Communication & Society* 18 (2): 121–38. https://doi.org/10.1080/1369118X.2014.940365.

Ruckenstein, Minna and Linda Lisa Maria Turunen. 2020. 'Re-humanizing the Platform: Content Moderators and the Logic of Care'. *New Media & Society 22 (6): 1026-1042. doi: 10.1177/1461444819875990.*

Santana, Arthur D. 2014. 'Virtuous or Vitriolic: The Effect of Anonymity on Civility in Online Newspaper Reader Comment Boards'. *Journalism Practice* 8 (1): 18–33. https://doi.org/10.1080/17512786.2013.813194.

Schwartz, Matthias. 2008. 'The Trolls Among Us'. *The New York Times*, August 3.

Slashdot. n.d. Frequently Asked Questions. Retrieved January 31, 2021 from https://slashdot.org/faq.

Suzor, Nicolas, Tess Van Geelen, and Sarah Myers West. 2018. 'Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda'. *International Communication Gazette* 80 (4): 385–400.

Suzor, Nicolas. 2020. 'Understanding Content Moderation Systems: New Methods to Understand Internet Governance at Scale, over Time, and across Platforms'. In *Computational Legal Studies*, edited by Ryan Whalen, 166–89. Edward Elgar Publishing. https://doi.org/10.4337/9781788977456.00013

The DIASPORA* project. n.d.. FAQ for users. Retrieved January 28, 2021 from https://wiki.diasporafoundation.org/FAQ_for_users

The Verge. 2021. Who decides what stays on the Internet? Regulation expert Daphne Keller on where moderation goes after banning Trump. By Nilay Patel. Jan 12, 2021. Retrieved January 15, 2021 from https://www.theverge.com/22225238/trump-social-media-ban-platform-moderation-tech-regulation-daphne-keller-interview

Towne, W. Ben, and James D. Herbsleb. 2012. 'Design Considerations for Online Deliberation Systems'. *Journal of Information Technology & Politics* 9 (1): 97–115. https://doi.org/10.1080/19331681.2011.637711

Townsend, Ellen, Emma Nielsen, Rosie Allister, and Sarah A. Cassidy. 2020. 'Key Ethical Questions for Research during the COVID-19 Pandemic'. *The Lancet Psychiatry* 7(5):381–83. doi: 10.1016/S2215-0366(20)30150-4.

Wagner, Ben. 2016. *Global Free Expression: Governing the Boundaries of Internet Content.* Cham, Switzerland: Springer International Publishing.

Wagner, Ben. 2016. 'Algorithmic Regulation and the Global Default: Shifting Norms in Internet Technology'. *Nordic Journal of Applied Ethics* 10 (1): 5–13.

Wagner, Ben. 2018. 'Free Expression? – Dominant Information Intermediaries as Arbiters of Internet Speech'. In *Digital Dominance: Implications and Risks*, edited by M. Moore and D. Tambini. Oxford: Oxford University Press.

Wagner, Ben, and Carolina Ferro. 2020. *Governance of Digitalization in Europe: A Contribution to the Exploration Shaping Digital Policy - Towards a Fair Digital Society?* Gütersloh, Germany: Bertelsmann Stiftung.

Wagner, Ben, Johanne Kübler, Lubos Kuklis, and Carolina Ferro. 2021. *Auditing Big Tech: Combating Disinformation with Reliable Transparency.* Tallinn, Estonia: Enabling Digital Rights and Governance & Omidyar Network.

Wagner, Ben, Krisztina Rozgonyi, Marie-Therese Sekwenz, Jatinder Singh, and Jennifer Cobbe. 2020. 'Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act'. Barcelona, Spain.

Wagner, Ben and Marina Kubina. 2021. Ergebnisse des Forschungsprojekts zur Stärkung der Diskussionskultur. in Der Standard. Retrieved February 21, 2021 from https://www.derstandard.at/story/2000124046106/ergebnisse-des-forschungsprojekts-zur-staerkung-der-diskussionskultur

Wang, Weiquan, and Izak Benbasat. 2007. 'Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs'. Journal of Management Information Systems 23 (4): 217–46.

Wikimedia. 2021. List of Wikipedias by language group. Retrieved February 1, 2021 from https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_language_group#Germanic_(15,921,922_articles,_179,418_active_accounts).

Wikipedia. 2020a. Wikipedia: Neutral Point of View. Retrieved February 2, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

Wikipedia. 2020b. Wikipedia: No personal attacks. Retrieved February 2, 2021 from https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks

Wikipedia. 2021a. Wikipedia: About. Retrieved January 30, 2021 from https://en.wikipedia.org/wiki/Wikipedia:About

Wikipedia. 2021b. Wikipedia: Administrators. Retrieved January 30, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Administrators

Wikipedia. 2021c. Wikipedia: Administrator intervention against vandalism. Retrieved January 30, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Administrator_intervention_against_vandalism

Wikipedia. 2021d. Wikipedia: Five pillars. Retrieved January 30, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Five_pillars

Wikipedia. 2021e. Wikipedia: Policies and guidelines. Retrieved January 30, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines.

Wikipedia. 2021f. Wikipedia: Wikipedians. Retrieved January 31, 2021 from https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedians&oldid=1004136172

Wikipedia. 2021g. Statistics. Retrieved January 31, 2021 from https://en.wikipedia.org/wiki/Special:Statistics

Wulczyn, Ellery, Thain Nithum and Dixon, Lucas. 2017. ‘Ex Machina: Personal Attacks Seen at Scale’. International World Wide Web Conference April 3–7, 2017, Perth, Australia, doi: 10.1145/3038912.3052591.

Yin, Robert K. 2018. *Case Study Research: Design and Methods*. Sixth edition. Los Angeles, CA: SAGE Publications Inc.

Zignani, Matteo, Christian Quadri, Alessia Galdeman, Sabrina Gaito, and Gian Paolo Rossi. 2019. ‘Mastodon Content Warnings: Inappropriate Contents in a Micro-Blogging Platform’. Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019). doi: 10.7910/DVN/R1HKVS.

Zulli, Diana, Miao Liu, and Robert Gehl. 2020. ‘Rethinking the “Social” in “Social Media”: Insights into Topology, Abstraction, and Scale on the Mastodon Social Network’. *New Media & Society* 22 (7): 1188–1205. doi: 10.1177/1461444820912533.

Abbreviations

CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
DSA	Digital Services Act
EU	European Union
FAQ	Frequently Asked Questions
ID	Identity
IP	Internet Protocol
NSFW	Not Safe for Work
OFAI	Austrian Research Institute for Artificial Intelligence
ToS	Terms of Service

Annexes

I. Semi-structured interview questionnaire

1. What is your role in relation to content moderation? How many years of experience do you have in the field?
2. What alternatives to deletion are you aware of for content moderation?
3. How effective are these alternatives in achieving their goals at content moderation?
4. Are there any unintended consequences of using these alternative content moderation techniques?
5. Do you have any empirical data on these alternative content moderation techniques that you would be able to share with us?
6. Do you think automated approaches to content moderation can be successful?
7. Are there any specific challenges with automated content moderation that you think we should consider?
8. Is there anything else important that you want to tell us?

II. Lists of interviews

	NAME OF INTERVIEWEE	DATE OF INTERVIEW	AFFILIATION
1	Marina Kubina & Christian Burger	16.12.2020	Der Standard
2	Jonne Haß	23.12.2020	diaspora*
3	Dennis Schubert	29.12.2020	diaspora*
4	Eric Goldman	06.01.2021	Santa Clara University School of Law
5	Caroline Sinders	11.01.2021	Weizenbaum Institute
6	Rob Malda	11.01.2021	(formerly) slashdot
7	Joseph Seering	13.01.2021	Carnegie Mellon University
8	Joan Barata	18.01.2021	Stanford Law School
9	Leighanna Mixter	19.01.2021	Wikimedia Foundation
10	Owen Bennett	20.01.2021	Mozilla Corporation

11	Giovanni Di Gregorio	26.01.2021	Bocconi University
12	Claudia Müller Birn	26.01.2021	Freie Universität Berlin, Wikidata
13	Kerry Kent	03.02.2021	change.org
14	Magdalena Piech	03.02.2021	Allegro
15	Leah Oswald	11.02.2021	mastodon
16	Niklas Henckell	25.02.2021	Jodel

III. Summary of interviews

1. Marina Kubina & Christian Burger, Der Standard

The meeting took place with Christian Burger, the head of community management at der Standard, and Marina Kubina, a community manager. Both Marina and Christian strongly emphasised their positive and constructive approach to content moderation. From their perspective, the main task of moderators is to promote positive interactions and prevent discussions from getting out of hand before something needs to be deleted. This approach is reflected in both their community-led perspective and their emphasis on the rights of their community to free expression.

All their approaches to forum design stem from this perspective, focusing on promoting positive conversations and community cohesion. They frequently collaborate with academics and other independent research institutes, as they believe that by themselves, they do not have sufficient resources to implement these kinds of design shifts. They also believe that their approach to content moderation is neither well understood nor sufficiently represented in public debates. As such, Christian Burger is writing a book on their approach to content moderation, which he hopes will shed further light on the matter.

2. Jonne Haß, diaspora*

Jonne Haß does content moderation for several small online communities, especially diaspora* and Freenode, with about 10 years of experience.

Haß reported that the deletion of content is somewhat effective in smaller communities. Often, however, communication suffices to ensure a civil discourse. The following alternatives to content deletion were suggested: Hiding content for all members but the author and moderators, banning short- or long-term users, and giving users the ability to moderate the content themselves. The latter proves to be essential and effective for the diaspora* community. The delegation of moderation to users can also cause disputes and impede discussions about divisive issues. Delegating moderation should not be the sole method of content moderation.

Automated content moderation can be valuable and effective if it is based on manually configured rules. Overall, human judgment is preferred to ensure accountability due to the higher risk of creating echo chambers and enforcing taboos when using automated content moderation.

3. Dennis Schubert, diaspora*

Dennis Schubert is a project manager at diaspora* and runs Geraspora, a diaspora* node with about 10 years of experience.

In the participant's experience, deplatforming and deleting content works, as bad actors stop posting if their contributions are deleted. As an alternative, a reputation-based content moderation system inspired by the Matrix project is currently investigated. Furthermore, user control mechanisms are improved to increase users' interaction possibilities. These measures effectively decrease targeted harassment campaigns, but they cannot control the spread of misinformation. User experience regression is a possible downside of such alternative content moderation tools.

Automated content moderation is effective only in combination with human reviewers. Moreover, any form of automated content moderation can be subject to abuse and might result in false positives, leading to user frustration.

4. Eric Goldman, Santa Clara University School of Law

Eric Goldman is a law professor at Santa Clara University School of Law and an expert on internet law with more than 20 years of experience.

Concerning content moderation, Goldman emphasised that setting the baseline at zero harm is unrealistic, as harm happens online and offline.

In terms of specific measures, Goldman sees two possibilities: messages are pre-screened and not published if deemed unfitting, which is unrealistic for platforms such as Twitter, of which a large part is happening in real-time. The other more likely possibility is the restriction of Twitter to brands or people who have a built-in audience they will bring to the platform. He considered the YouTube and Facebook platforms to be doing interesting things regarding content moderation, especially the latter, which implemented more transparency measures than others. Overall, Goldman affirmed that all platforms are doing interesting things regarding content moderation in their environment. Moreover, he warned that a one-size-fits-all approach would not be feasible, as communities are sensitive to different issues. Although some regulations might solve one problem, they might cause several challenges for other communities.

5. Caroline Sinders, Weizenbaum Institute

Caroline Sinders analyses content moderation in the context of research on online harassment and advocates for content moderators' safety with about 7 years of experience.

Sinders states that taking down content can be effective in some cases, especially if the content endangers people, spreads disinformation, and causes harm. Moreover, the participant advocates for storing content where the wider public cannot access it. Certain tools and procedures should be included prior to the question of deletion of content, such as the possibility of saving a draft on the platform, training content moderators to deal adequately with harassment, and identifying if something is harassment.

Especially considering harassment, Sinders does not believe that automated content moderation could be successful. Harassment, toxicity, and abuse are almost impossible to filter out using an AI filter for sentiment analysis. Due to its cultural dependence and use of microaggressions, harassment is difficult to identify. Moreover, in cases of stalking, for instance, it is difficult to prove and report. Therefore, Sinders argues that automated content moderation might create more difficulties than it would solve.

6. Rob Malda (formerly), slashdot

Rob Malda created and led the platform slashdot for 14 years, one of the first larger-scale moderation systems, which was entirely community driven and had a policy of not deleting content.

Deletion can, according to Malda, be successful sometimes. However, it depends on the platform's scale, especially as trolls are reincentivized to act out if their content is deleted, which makes the problem worse. The system developed for slashdot provided an alternative to deletion by using a user-driven scoring system, and the content marked down by other users was essentially hidden. However, many instances of downvoting one user could inspire anger and frustration, which might incentivise them to act out further. Moreover, the platform used a meta-moderation system, meaning that users would get points to moderate other moderation. To impede bad moderation, people who spend the most time posting about a certain issue could not moderate that particular content. Furthermore, after a user upvotes or downvotes content, a jury of other users determines whether that moderator has done so fairly.

7. Joseph Seering, Carnegie Mellon University

Joseph Seering has researched content moderation for about 5 years, doing empirical work on different platforms and focusing on platforms that allow volunteer moderation. The effectiveness of deleting content depends, according to Seering, on the situation, as deleting the initial problematic piece of content can be effective at discouraging additional problematic behaviours. Deletion is especially effective if human intervention in the form of conversation is not fruitful, as in the case of users without intentions to contribute positively or bots. Alternatives can be efficient, each depending on the reasons for rule violations; for example, if a user can be educated, they might contribute positively in the future. Seeking conversation and solving disputes, however, requires more labour and is usually done by volunteers. Automated content moderation can be effective if it supports content moderators and filters out problematic content, such as pornography, to protect moderators. Granting users more autonomy to manage their spaces might be the best way to establish effective content moderation on a large scale.

8. Joan Barata, Stanford Law School

Joan Barata is a scholar with the Centre for Internet and Society at Stanford Law School, studying intermediary liability regulations for content moderation, with 10 years of experience in that field.

Barata posits that some forms of content require deletion, probably even deleting an account in extreme cases, and it has to be agreed on where that sort of extreme measure is necessary. Importantly, a choice between content moderation resulting from a legal framework imposed by states or the result of the initiative of platforms has to be made. Barata argues that the former is dangerous for freedom of expression. Moreover, infrastructure providers are good moderators of the platforms' content moderation.

Alternative forms of content moderation mentioned are mostly interstitial, which entails hiding content, making it less visible, or flagging it as inappropriate. In general, content moderation is unavoidable, and the effectiveness of the adopted measures is key. Infrastructure providers are not good moderators of platforms' content moderation. They hold great power vis-a-vis platforms, but this capacity to intervene is neither good nor adequate in terms of freedom of expression. If these measures are badly implemented, errors happen. In that regard, a proper impact assessment is crucial. Sometimes, taking down content deprives people of the content that they have a right to know and, therefore, impedes the plurality and diversity of opinions, ultimately representing a false picture of reality. Automated systems are needed and work well if the context of the content is not relevant.

9. Leighanna Mixter, Wikimedia Foundation

Leighanna Mixter is a senior legal counsel for the Wikimedia Foundation with about 5 years of experience. Mixter is involved in the transparency report by the Wikimedia Foundation, public policy creation in the United States, and their anti-censorship portfolio.

According to Mixter, the effectiveness of deleting content depends on the circumstances, as some form of content, which is illegal in almost every context, needs to be removed. For other types of content, different forms of moderation are more viable to guarantee freedom of expression. The content might be flagged, provided with additional information, or hidden for some time or regions. If a platform aims to ensure that nobody sees explicit material, automated content moderation might be a great way, even as it generates false positives. For each goal, however, the right type of content moderation needs to be found.

Mixter suggested that automated systems can be very effective; there might be issues surrounding biases of training datasets and lack of context awareness, for example, regarding copyright. The awareness of the automated systems' limits is important to consider. Overall, the human element is important, especially the diversity and representativeness of human moderators, which is valuable to reduce false positives. In this regard, it is important for regulators to leave room for community-led models of content moderation in all respects, as community governance is a more suitable or desirable model for some platforms. This includes Wikipedia, where Wikimedia volunteer editors take the lead role in content moderation decisions; however, moderation also includes community moderators on small message boards, Reddit's subreddits, Facebook groups, Discord channels, and more.

10. Owen Bennett, Mozilla Corporation

Owen Bennett works for the Mozilla Corporation, focusing on platform regulation issues, with about 8 years of experience.

According to Bennett, the effectiveness of content deletion depends on the context. Especially in Europe, where certain forms of speech are prohibited, deletion will always be a form of content moderation. Using removal as the primary method of content moderation needs to change rapidly. Other forms of content moderation, such as tweaking recommender systems, downranking, or deranking problematic content are available. These, however, are decided by the platforms themselves, and content moderation is done voluntarily. Alternative forms of content moderation and governance, which focus on the presentation of content and how it shapes user experience, are important. The data about content moderation done by big platforms, such as Facebook, YouTube, and Twitter, are not available for researchers or policymakers. Therefore, the effectiveness of these alternative forms of content moderation is not known.

Some degree of automation in content moderation is likely necessary for larger platforms, given the sheer volume of content on their services. Automation can be effective in some contexts, for example, most notably for the detection of child sexual abuse material. However, such software is proprietary, so smaller platforms might not have the resources to use them and need to rely on 3rd party junk solutions that are ineffective and might be harmful to fundamental rights.

11. Giovanni De Gregorio, Bocconi University

Giovanni De Gregorio is a researcher working on constitutional law and technology, focusing on social media, content moderation, and its challenges for free speech, with about three years of experience in that field.

According to De Gregorio, deletion is the quickest and most reactive approach to dealing with objectionable content. In some cases, removal can be effective. Removal of content is a business decision made by platforms that is influenced by interest in preserving advertising revenues. We have seen how the platform can also profit from the spread of content that captures a high degree of engagement, such as hate speech and disinformation. Top-down and bottom-up pressures might, however, influence these decisions and make platforms to review their policies, as in the case of Myanmar. Nonetheless, social media also has other remedies aside from removal. Removal executed by users might be possible but depends on the power of the community. Therefore, automated approaches to content moderation are necessary, as users and human moderators cannot cope with all the content on big platforms. Among alternative remedies, it would be possible to mitigate the risk of removal ex-ante by intervening on algorithms or increasing the degree of public and/or independent audit in the field of content moderation. Moreover, a new form of media pluralism could be designed online. In that regard, the logic of content moderation should not be driven just by advertising but also by public interest logic. For instance, taking into account minority opinions could be a step towards promoting more diversity in content moderation.

12. Claudia Müller-Birn, Free University Berlin, Wikipedia

Claudia Müller-Birn is a professor for Human-Centred Computing at the Free University Berlin and has worked on Wikipedia projects for about 15 years.

According to Müller-Birn, deleting content is not necessarily the best strategy, especially as the content is already in the system. Moreover, a private company should not decide what is deleted or what users can see. Wikipedia has a very successful system of organising content moderation. Legislation alone cannot solve the problems of content moderation, and any content moderation strategy should include the users. Concerning platforms, two forms of governance can be distinguished: algorithmic governance and juridical governance. Both need to be taken into account.

A general problem with content moderation concerns moving targets. The rules formalised in software cannot be changed or circumvented, but the prioritisation of rules in a society changes. Automated approaches to content moderation can be successful in some areas. What the automated system does, how it is trained, and how good it works are crucial issues. Thus, compliance is important, and controlling mechanisms should be established.

13. Kerry Kent, change.org

Kerry Kent has been the Global Head of Policy at change.org, working on content moderation for about 20 years.

On the petition platform, change.org, users are allowed to post any petitions that are important to them. They agree to guidelines such as no hate speech, no shocking images, no harmful language, and no misinformation that is likely to cause harm. Kent does not believe deletion to be an effective form of content moderation, and the platform tries to avoid it, except for cases of illegal content or a clear breach of guidelines. If the content evaluation is unclear, users might need to edit their submissions or provide additional evidence for their claims. Generally, petitions that are considered part of the usual discourse can remain on the platform.

Automated systems are in place to filter out spam and explicit material. More specifically, post publication, the platform uses a search tool to scan content and filter out illegal content and spam. The filter picks up potentially abusive or bullying content as well as spam and illegal content. This content is sent for human review but remains on the platform until it has been verified to be contrary to the community guidelines. No content is pre-moderated, and content is reviewed if users flag it. Automated systems can be successful as they learn to handle increasingly nuanced data, and they can be useful to cover more ground. However, human moderators should be involved in the process. On the downside, automated systems might flag content that does not violate the platform's guidelines. Moreover, automated systems cannot adapt flexibly to events as they unfold, which is sometimes a challenge.

14. Magdalena Piech, Allegro

Magdalena Piech is the head of regulatory affairs for Allegro, an e-commerce platform based in Poland. Allegro is the largest online marketplace in Poland, with about 99% of sellers being third party platforms. The platform mainly provides the technological infrastructure for transactions to be carried out smoothly. Sellers register on the platform, and subsequently Allegro verifies their background information. The platform currently has 163 million active offers. The platform has a notice and take-down system. Any user can report a product and choose reasons from a dropdown menu, for instance, because it is illegal, unsafe, or misleading. Subsequently, the security team verifies the report and takes down the product if necessary. If products are flagged, sellers are warned before the product is evaluated and possibly taken down. Users can rate products after transactions take place.

Illegal or unsafe products are blocked, and to ensure that the decision is justified, Allegro cooperates with authorities and NGOs. Moreover, additional measures against scams and the distribution of misinformation or content promoting hate speech are in place. Automated systems are used in some cases, but only in combination with human operators who verify the system's decisions.

15. Leah Oswald, mastodon

Leah Oswald has been an administrator of the mastodon instance chaos.social since 2017.

Oswald states that deletion can be effective as a last resort for harmful content. However, on mastodon, administrators try to talk to users first and only silence and delete accounts or instances if a violation happens multiple times and if the user is a troll or a spammer. The participant emphasises that moderation should be transparent, understandable, and balanced. Otherwise, the moderators might lose the users' trust. Moreover, rules must be established that every user has to accept before using a platform, which facilitates the justification of decisions.

Automated approaches to content moderation are, according to Oswald, successful only for limited cases, such as spam, due to the importance of the context of the content. Moreover, every algorithm is biased and cannot understand all the possible ways in which people interact and interpret subtexts.

16. Niklas Henckell, Jodel

Niklas Henckel is Head of Expansion of Jodel, a social media app for anonymous communication with people in the immediate surroundings of the user. Henckel is one of Jodel's founders and used to be head of community with six years of experience in content governance.

Henckel points out that the deletion of content can be effective if connected to other rules and curation mechanisms, as it is necessary to have multiple safety measures to be effective and to educate abusers through careful punishments or, if necessary, remove them from the community. Alternatives to deletion that might be effective are shadow banning—hiding content to everybody except the poster—having separate rooms for adults or other content that might be inappropriate for some groups and making it known that authorities are informed of any illegal content. Limiting access to content with many negative ratings is also effective. The best strategy is to seek conversation and explain breaches of conduct to people, which, however, requires many resources.

Automated approaches to content moderation can be successful in unambiguous cases, such as paedophile content. These types of content, however, are not the majority of the content that is posing problems. Especially regarding text moderation, it is very difficult to build good and effective models. Moreover, regulations need to take into account that smaller platforms do not have the resources to develop such effective automated systems.

© Picture credits:
Markus Spiske - Unsplash(Cover)

THEGREENS/EFA
in the European Parliament

60 rue Wiertz/Wiertzstraat 60
1047 Brussels, Belgium
www.greens-efa.eu
contactgreens@ep.europa.eu