● TDP

Transformationen des Populären.
Working Paper Series des SFB 1472

25.05.2022

Working Paper 3

From healthy communities to toxic debates: Disqus' changing ideas about comment moderation.

Johannes Paßmann
Anne Helmond
Robert Jansma

Abstract

This article examines how the commenting platform Disqus changed the way it speaks about commenting and moderation over time. To examine this evolving self-presentation, the Internet Archive Wayback Machine was used to analyse the company's website and blog between 2007 and 2021. By combining interpretative close-reading approaches with computerised distant-reading procedures, it examined how Disqus tried to advance online discussion and dealt with moderation over time. It was found that in the mid-2000s, commenting systems were supposed to help filter and surface valuable contributions to public discourse, while ten years later their focus had shifted to the proclaimed goal of protecting public discourse from contamination with potentially harmful ("toxic") communication. To achieve this, the company developed new tools and features to facilitate and semi-automate this active and interventive form of moderation. This rise of platform interventionism was fostered by a turn towards semantics of urgency in the company's language to legitimise its actions.

Keywords: Disqus, commenting systems, online comments, moderation, toxic, web archives

Introduction

Discourse around online commenting has seen significant changes in the past few decades. Earlier, its proponents were not concerned with "combating hate speech" or "toxic communication", but rather with questions such as: Can comments be considered as texts worth reading? How can commenting technologies engage more users? And, perhaps most importantly, how can commenting systems improve public deliberation?

Research on comments points to a history of (increasing) problematisation. As the scale of online commenting grew — along with the number of technologies enabling website owners to implement commenting systems — the problems with commenting scaled as well. While early issues around comments related mostly to spam, they are now seen as being central to strategic misinformation, targeted harassment, and trolling campaigns.[1] This shift shows that while in the early 2000s comments were commended for their role in providing a space with a new "deliberative democratic potential",[2] they have been de-valorised in the past decade and now belong to the "bottom half of the web".[3] More generally, the online media landscape has been diagnosed as being in a state of "network climate change", referring to a systemic transformation of the web caused by "pollution" emitted by all kinds of actors.[4]

This article examines this shift by exploring how Disqus, a central producer of online commenting software, has changed its language about its technology. Disqus, founded in 2007, became a popular commenting system as it enabled any webmaster to add a commenting system to their website by simply embedding a snippet of code or by installing a plugin. Since Disqus provides the tools for both running and managing a commenting section, it quickly became a key technology used by major (news) websites and (technology) blogs such as CNN, Wired, Russia Today, and Breitbart. Currently, Disqus is one of the oldest and most widely used third-party commenting systems on the web.

1 Charlie Warzel, 'How The Alt-Right Manipulates The Internet's Biggest Commenting Platform', BuzzFeed, 2018 <https://www.buzzfeednews.com/article/charliewarzel/how-the-alt-right-manipulates-disqus-comment-threads> [accessed 14 September 2021].

2 Luke Collins and Brigitte Nerlich, 'Examining User Comments for Deliberative Democracy: A Corpus-Driven Analysis of the Climate Change Debate Online', *Environmental Communication*, 9.2 (2015), 189–207.

3 Joseph M. Reagle, *Reading the Comments: Likers, Haters, and Manipulators at the Bottom of the Web* (MIT Press, 2016).

4 Whitney Phillips and Ryan M. Milner, *You Are Here: A Field Guide for Navigating Polarized Speech, Conspiracy Theories, and Our Polluted Media Landscape* (The MIT Press, 2021).

5 BuiltWith, 'Comment System Usage Distribution in the Top 1 Million Sites', 2021 <https://trends.builtwith.com/widgets/comment-system> [accessed 9 September 2021]; Disqus, 'About Disqus', 2021 <https://disqus.com/company/> [accessed 9 September 2021].

6 Rebecca Blood, 'How Blogging Software Reshapes the Online Community', *Communications of the ACM*, 47.12 (2004), 53–55; Axel Bruns, Gatewatching: Collaborative Online News Production (Peter Lang, 2005).

7 Blood.

8 Anat Ben-David and Oren Soffer, 'User Comments across Platforms and Journalistic Genres', *Information, Communication & Society*, 22.12 (2019), 1810–29; Ian Rowe, 'Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms', *Journal of Broadcasting & Electronic Media*, 59.4 (2015), 539–55.

9 Bruns; Thomas B. Ksiazek and Nina Springer, *User Comments and Moderation in Digital Journalism: Disruptive Engagement* (Routledge, 2020).

10 Ksiazek and Springer.

11 Lena Frischlich, Svenja Boberg, and Thorsten Quandt, 'Comment Sections as Targets of Dark Participation? Journalists' Evaluation and Moderation of Deviant User Comments', *Journalism Studies*, 20.14 (2019), 2014–33.

12 Renee Barnes, *Uncovering Online Commenting Culture: Trolls, Fanboys and Lurkers* (Palgrave Macmillan, 2018); Ben-David and Soffer; Rowe.

13 Reagle.

14 Frischlich, Boberg, and Quandt; Ksiazek and Springer; Reagle.

15 J David Wolfgang, 'Taming the "Trolls": How Journalists Negotiate the Boundaries of Journalism and Online Comments', *Journalism*, 22.1 (2021), 139–56.

As of September 2021, 1.64% of the top 1 million websites were using Disqus, with over 50 million comments per month.[5]

The self-presentation of a software company has strategic motives and provides insights into the discursive construction of a technology and its practices. Thus, rather than analysing the writing of comments, their moderation, or maintenance of commenting technologies, we study the evolution of discursivation of those technologies and practices from the perspective of their producer. We argue that the evolving self-presentation of Disqus has reasons, drivers, and effects that are relevant for understanding the transformations of commenting specifically and digital cultures in general.

Studying Online Commenting (Systems)

Discourses around online commenting have changed considerably over the last three decades, and two main perspectives can be differentiated in the literature: the participation and democratisation enabled by commenting, largely referring to end-user practices, and the subsequent moderation of comments, referring to the practices of website owners.

Participation and democratisation

Since the mid 1990s, comments have become central in online communication and blurred the boundaries and hierarchies between text producers (writers) and consumers (readers). Early accounts of online comments discussed their potential in democratising online publishing by offering users a novel mode of participation.[6] Consequently, user comments were considered a central and defining element of blogs. Along with giving blog readers a voice, they also enabled "cross-blog talk" and made the readers visible.[7] Online comments were conceptualised as novel facilitators of online dialogue with the potential "to foster public deliberation and civic discourse".[8] Furthermore, they could transform discursive hierarchies by turning commenters into active participants.[9] However, others found how this deliberative space could be "disrupted",[10] signalling how commenting later turned into a form of "dark participation" requiring moderation.[11]

Moderation

The different affordances of commenting technologies enable and constrain types of commenting practices and cultures, and their moderation.[12] For instance, commenting systems may assist in moderation and help webmasters "fortify" their websites to minimise comment abuse.[13] The rise of comment moderation and its (semi-)automated technologies is frequently connected with increasingly disruptive usage practices. This new discourse of disruption, dark participation, and fortification in commenting literature[14] is also visible in the language describing the person of the commenter who is seen "as a threat" that requires "taming".[15]

16 Jin Woo Kim and others, 'The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity', *Journal of Communication*, 2021.

17 Nicholas Diakopoulos, *Automating the News: How Algorithms Are Rewriting the Media, Automating the News* (Harvard University Press, 2019), p. 73 .

18 Alexander Halavais, 'Do Dugg Diggers Digg Diligently?', *Information, Communication & Society*, 12.3 (2009), 444–59 .

19 Michael Stevenson, 'Slashdot, Open News and Informated Media: Exploring the Intersection of Imagined Futures and Web Publishing Technology', in *New Media, Old Media: A History and Theory Reader*, ed. by Wendy Hui Kyong Chun, Anna Watkins Fisher, and Thomas Keenan (Routledge, 2016), pp. 616–30.

20 Stevenson, p. 622.

21 Ronja Trischler and Beatrice Perera, 'Digitale Rituale Moderieren' (unpublished Paper presented at the Workshop Digitale Rituale, Frankfurt a.M.).

22 Albert Ziegler, 'Wer Moderieren Will, Muß Maß Nehmen Und Maß Geben.', *Gruppendynamik-Zeitschrift Für Angewandte Sozialpsychologie*, 23.3 (1992), 215–36.

23 Josef W. Seifert, 'Moderieren', in *An-Gewandte Sozialpsychologie: Das Praxishandbuch*, ed. by A. E. Auhagen and H. W. Bierhoff (Beltz PVU, 2003), pp. 75–87.

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

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

Moreover, the discourse around comments now frequently characterises problematic comments as "toxic" or a type of "uncivil communication".[16] This naturalistic (partially pathological) discourse of toxic comments, as something that needs to be cured, is particularly relevant for understanding how the semantics of online commenting have changed during the last decade.

As Diakopoulos has observed, the overwhelming number of comments has led major news websites to deploy machine-learning models and filtering algorithms to automate journalists' commenting moderation decisions.[17] In this novel socio-technical assemblage, algorithms assign scores to comments or commentators for (semi-)automated filtering and decision-making processes. Other users may also contribute to this collaborative, crowdsourced moderation process by assigning scores to comments through upvoting, downvoting, liking, disliking, or other evaluative criteria and related metrics.[18] Importantly, Stevenson describes how the rise of content management systems and cookies during the late 1990s meant that comments could now be stored into a database and connected to a user. This development helped turn comments and commenters into "data-rich objects" that could be assigned scores, points, or a reputation score.[19] These calculated metrics and evaluative criteria helped with "scaling up conversation", and actively shaped it by filtering out the "noise" to surface "quality" comments.[20]

Finally, more general studies on online moderation show how it has always been associated with active intervention in a potentially adversarial situation.[21] Already the etymology of moderation implies an interventive activity as it has its roots in the Latin "moderatus", meaning modest, tempered, calm, or judicious,[22] thereby emphasising something that needs to be restrained or kept within bounds. This counts also for the social psychology of moderation where Seifert fleshes out that the core function of moderation is to mediate conflict and to reconcile conflicting positions.[23] On the web, this active intervention is performed by platforms, such as social media and commenting platforms, who have centrally positioned themselves as mediators or even "custodians" of the public debate.[24] Furthermore, as Roberts has demonstrated,[25] a whole commercial content moderation industry has emerged for dealing with comments sections and other online content.

26 Sungho Shin et al., 'Analytics on Online Discussion and Commenting Services', in *Human Interface and the Management of Information. Information and Interaction Design.* (Springer, 2013), pp. 250–58 .

27 Hasan Al Maruf et al., 'Human Behaviour in Different Social Medias: A Case Study of Twitter and Disqus' (presented at the IEEE/ACM International Conference 2015, 2015) .

28 Emma Goodman, 'Online Comment Moderation: Emerging Best Practices' (World Association of Newspapers, 2013) .

29 Eli Omernick and Sara Owsley Sood, 'The Impact of Anonymity in Online Communities', in *2013 International Conference on Social Computing*, 2013, pp. 526–35 .

30 Al Maruf et al.; Daniel de Zeeuw and others, 'Tracing Normiefication: A Cross-Platform Analysis of the QAnon Conspiracy Theory', *First Monday*, 2020.

31 Johannes Paßmann, 'Medien-theoretisches Sampling. Digital Methods als Teil qualitativer Methoden', *Zeitschrift für Medienwissenschaft*, 13.2 (2021), 128–40 .

32 John M. Swales and Priscilla S. Rogers, 'Discourse and the Projection of Corporate Culture: The Mission Statement', *Discourse & Society*, 6.2 (1995), 223–42 .

33 Shelton Weech, 'Analyzing Organizational Values Over Time: A Methodological Case Study', in *Proceedings of the 38th ACM International Conference on Design of Communication,* SIGDOC '20 (Association for Computing Machinery, 2020), pp. 1–6 .

Lastly, few case studies have been solely dedicated to this company and its technology. Most studies either outline Disqus' general features and affordances,[26] compare them to other commenting systems and social media platforms,[27] describe how they assist editors and community managers of news organisations in their work practices,[28] or examine the effects of Disqus' affordances for anonymity and pseudonymity on the quality of comments and the conversation.[29] Finally, researchers have used the Disqus application programming interface (API) to sample comments and to characterise their contents, their spread, or the behaviour of commenters.[30]

What we contribute to this wide-ranging body of literature on the practices, systems, and moderation of commenting, is a detailed study on Disqus's discursive positioning of its commenting technology. As such, we do not focus on the effects of the technology on its users, commenting communities, or quality or quantity of conversations, but rather on how a company that has built a key and widely used commenting system has shaped its ideas about comment moderation over time.

Method and Materials

Internet platforms — be they social media or commenting platforms — produce a wealth of information and documentation about their companies' business operations and technologies. These (archived) materials can be employed to examine how they discursively position themselves by analysing how they describe their users, use cases, and business models of their technologies and their role in digital society more generally.[31] Within organisational studies, company vision and mission statements have been used as central documents for understanding how companies "project their corporate philosophy" onto the world.[32] Or, using web archive data from company homepages, for understanding how the "organization values" of companies have evolved over time.[33]

34 Gillespie, p. 47.

35 Richard Rogers, 'Debanalizing Twitter: The Transformation of an Object of Study', in *Proceedings of the 5th Annual ACM Web Science Conference* (ACM, 2013), pp. 356–65; José Van Dijck, *The Culture of Connectivity: A Critical History of Social Media* (Oxford University Press, 2013); Anna Lauren Hoffmann, Nicholas Proferes, and Michael Zimmer, '"Making the World More Open and Connected": Mark Zuckerberg and the Discursive Construction of Facebook and Its Users', *New Media & Society*, 20.1 (2018), 199–218 .

36 Jean Burgess, 'From "Broadcast Yourself" to "Follow Your Interests": Making over Social Media', *International Journal of Cultural Studies*, 18.3 (2015), 281–85

37 Paßmann.

38 Barney G. Glaser and Anselm Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research* (New York: Aldine, 1967).

39 Kathy Charmaz, *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis* (Thousand Oaks, CA: SAGE, 2006).

40 Paßmann.

41 Laura K. Nelson, 'Computational Grounded Theory: A Methodological Framework', *Sociological Methods & Research*, 49.1 (2020), 3–42.

42 Nelson, p. 5.

Within media studies, analysing Internet companies' (evolving) self-description is an established approach. Community guidelines of Internet platforms have been examined for doing "important discursive work" because they "articulate the 'ethos' of a site".[34] In their turn, company taglines reflect the suggested, imagined, and potential uses of a platform,[35] which makes them performative because usage practices can follow these imaginations.[36] At the same time, these representations are frequently a consequence of how platforms observe their users' practices. To examine this double function of a company's self-description as an effect and a driver of changing practices, we turn to the Internet Archive Wayback Machine (IAWM), to locate archived platform materials that enable "self-description histories" of web companies and their technologies.[37]

Sampling

To analyse Disqus' discursive positioning over time, we employ an approach oriented towards grounded theory methodology,[38] by sampling, coding, and analysing our data in iterative-cyclical procedures. In the sense of the constructivist variant of grounded theory,[39] we acknowledge the preconceptions shaping our methods, rather than suggesting a possibility to develop concepts "purely from data". As we are making use of digital research methods, it is important to underscore not only the active production of preconceptions, but also of our tools used to sample data within qualitative data analysis.[40] This required us to adapt our methods as a result of these methodological considerations.

Consequently, we did not only triangulate our data, but also our tools by combining interpretative close-reading approaches with computerised distant-reading procedures, as discussed in "computational grounded theory".[41] This method "combines expert human knowledge and skills at interpretation with the processing power and pattern recognition brought by computers".[42] Our method resembles computational grounded theory insofar that we combined computerised and non-computerised reading techniques in an iterative-cyclical, inductively hypothesis-generating procedure.

We began with a close reading of Disqus' front-page and its evolving taglines, to gain a general overview of changes in the company's history. For this purpose, we collected one archived snapshot from the IAWM of the Disqus homepage per year between 2007 and 2021. Here, our core assumption was that changes in how the company presents itself on its front-page and through its taglines may help to generate further, more specific hypotheses. During this initial analysis, it became evident how different ideas of enabling "better" discussions evolved over time.

In a second step, we sampled a full-text corpus of more specific data about the organisation's self-presentation from the company's blog, to specify how Disqus tried to improve online discussions (e.g., by introducing specific features) and as a reaction to what (e.g., as a result of the proliferation of trolling practices). Thus, we collected all the company's blog posts from both the live web and the IAWM between February 2007 and May 2021, retrieving 646 in total (see Table 1). While Disqus keeps a collection of its blog posts on blog.disqus.com, the blog's "self-archive" on the live web is not complete and only goes back to 2010. By going to the IAWM and following the links from the Disqus homepage to the company's blog, we found that

the URL of the main blog changed from blog.disqus.net (2007–2009) to blog.disqus.com (2010–ongoing). We collected these additional blog posts from blog.disqus.net from the IAWM. By following the links in the IAWM, we also discovered that Disqus operated multiple company blogs (business.disqus.com/blog, publishers.disqus.com/blog), and included these additional blog posts in our analysis.[43] Thus, we were able to reconstruct the blog's full archive by tracing it on the live web and in the web archive and triangulating these sources.

43 The additional posts found on http://code.disqus.com/ have not been included because they discuss the technical infrastructure of Disqus.

year	#
2007	31
2008	71
2009	78
2010	61
2011	23
2012	42
2013	39
2014	39
2015	25
2016	55
2017	44
2018	38
2019	46
2020	45
2021	9

Table 1. Blog posts per year

As this data corpus of 646 company blog posts was too big for close reading, we turned to computational content analysis to operationalise a distant reading approach based on keyword resonance. We started from the premise that changing notions of how Disqus enables online discussions and moderation are associated with the rise and fall of certain keywords. Consequently, we created a list of keywords (see Table 2) for probing in the text in multiple steps. We derived keywords from the literature and our domain knowledge on commenting and moderation and augmented this list in an iterative-cyclical procedure with keywords derived from our close reading and coding of the blog posts (see below). To make sure that we did not miss potentially relevant keywords with this inductive approach, we extracted the 1000 most used words from the blog posts and added newly detected relevant keywords to our list. Some words were stemmed to allow for capturing variations on nouns and conjugations of verbs. We searched through all blog posts for our keywords using regular expressions and counted the number of keyword occurrences per blog post.

Keywords	moderat (132), admin (67), approv (28), ban |bann (33), blacklist (5), block (31), catch (37), delet (36), filter (45), flag (30), identify (59), kick (27), remov (48), report (80), restrict (16), rule (32), timeout (7), whitelist (6), featured (45), reputation (24), trust (66), verified (11), harass (6), spam (63), toxic (9), troll (29), downvot (14), upvot (30).

Table 2. Moderation-related keywords

During the distant reading of all blog posts, it turned out that it was necessary to better understand the evolving context of the semantics of moderation, as the term invoked many different meanings. This demonstrates how our distant computational methods informed the need for an additional interpretative close reading of a selected set of posts relating to moderation. From the 646 blog posts, we created a sub-sample of 132 posts that contained the stem "moderat". This sub-sample was loaded into the qualitative coding software MAXQDA for further coding and analysis. Our coding, next to the augmentation of the keyword list above, resulted in a system with 18 categories containing 1444 codes, where the largest category was "moderation practices" with 424 codes. In addition, we performed a close reading on the posts containing moderation-related keywords to further understand the role of these metrics, features, and practices in relation to moderation (see Table 2).

Commenting Moderation Practices and their Discursivation

The analysis of Disqus' evolving front-page and tagline reveals a shift in how the company perceived the role of its commenting technology, intended uses, and envisioned users. In the early years, Disqus showcased its ambitions to improve online discourse by "Powering better discussion" (2007), describing its technology as "a powerful comment system that easily enhances" (2007) or "improves" (2008) the discussions on websites. Later, the focus shifted to connecting "conversations across the web" by enabling a form of "webwide discussion" (2008–2011). Next, it emphasised the technology's capacity to help build "active communities" from a website's commenters (2011–2015). This notion of the "community" remains very central between 2010–2015, as Disqus positions itself as a website and tool for discovering and building communities around conversations, even calling itself "The Web's Community of Communities" (2013–2015).

As of 2016, the company starts to address professional publishers as its central user group, instead of individual bloggers and website owners. Disqus is no longer just a tool for creating "communities" around a website's comments, but especially for building website "audiences" that can be monetised with ads (2016). In this development, Disqus becomes a tool that helps publishers to "increase engagement and build loyal audiences" (2017). Shortly after, the company was acquired by Zeta Global, a marketing technology company (DB-2017-12). In their acquisition blog post they discuss leveraging Zeta's expertise in artificial intelligence by using "machine learning to significantly reduce the burden of human moderation" (DB-2017-12). This signals the increasing role of moderation in managing comments and the need for more sophisticated technological solutions.

Whereas earlier taglines defined ideals for online commenting practices, the focus later shifted to the professional and technical solutions Disqus provides, with a new focus on (semi-) automation. Thus, the implicit reader of these taglines changes from (small) website owners and a general public interested in deliberation and community-building to a specific group of customers interested in efficient (tool-assisted) and monetizable commenting practices. Our keyword analysis supports this observation and shows how moderation progressively becomes an issue about how Disqus presents

itself, with a consolidation towards technical solutions for specific problems.

The rise of moderation

The number of blog posts with and without a moderation-related keyword can be seen in Figure 1. Three main periods can be differentiated. The first, from 2007 to 2010, shows a high level of blog activity, as the company blogs itself into existence, but only one fourth to one third of the posts contain a keyword related to moderation.
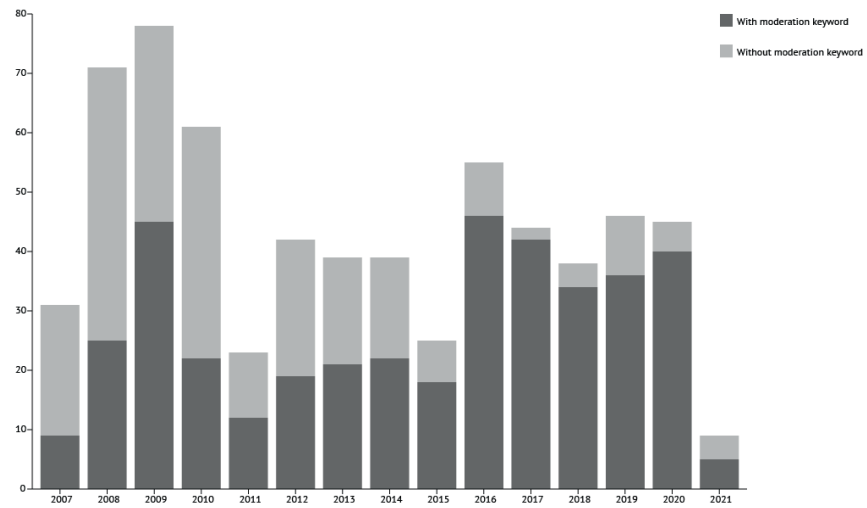

Figure 1. Blog posts with and without moderation keywords per year.

The second period, from 2011 to 2015, shows a decrease in the overall number of posts but moderation-related posts gain more prominence. The third period, from 2016 onwards, shows a slight increase in the total number of posts, but most of them mention moderation-related keywords. This indicates that when Disqus publicly writes about its technology, it seems to be increasingly concerned with issues pertaining to comment moderation. This may mean that Disqus is becoming further interested in helping their professional customers cope with their moderation issues rather than conveying the value of online commenting to a general public. Or, more generally, it signals that the issues with online commenting have evolved and become more prominent over time.

Keyword resonance in Disqus' company blog posts
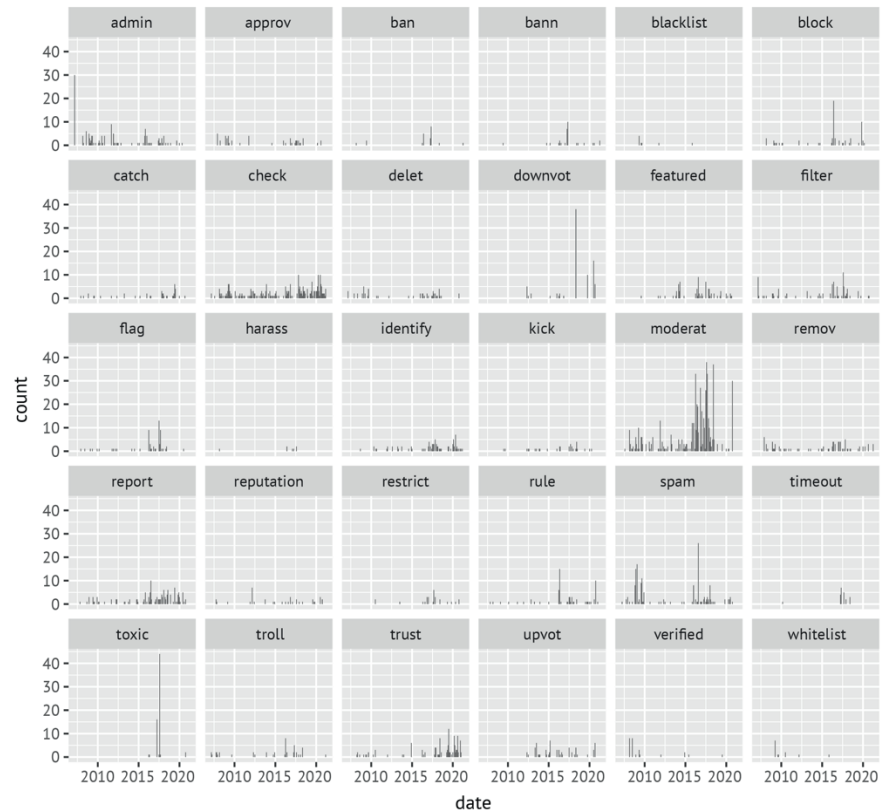
Moderation-related keywords

Figure 2: Moderation-related keywords in the blog posts per year. Figure created with the assistance of Martina Schories.

Figure 2 takes a closer look at which moderation-related keywords are mentioned in the blog posts. We roughly observe the same periods as before and detect an increase in moderation-related keywords as of 2016 (the year of the Brexit vote and the Trump election), including "moderat*" in general, but also more specific moderation-related practices such as "ban | bann", "block", "downvot*", "flag", "filter", and "report". These keywords signal a larger diversity or differentiation of Disqus' moderation practices. In 2016–2017, there is a spike in the keyword "toxic" which we will elaborate on in more detail below. After 2017, the general keyword "moderat*" diminishes, while more specific keywords such as "trust", "identify", and "timeout" gain more prominence. We read this as a sign of simultaneous consolidation and specification concerning the way Disqus thinks and speaks about its moderation features and practices. To better understand these evolving practices, we turn to our close reading of blog posts mentioning "moderat*" and other moderation-related keywords.

Overall, the blog posts demonstrate how the company developed different ideas about online conversations and moderation strategies over time, and discusses the changing role of metrics and features in their tools for moderation. A recurring issue in those posts was the role of popularity markers.

Popularity metrics assisting moderation

In August 2007, Disqus launched its "Beta" version and introduced new fea-tures to its commenting technology, including a follow function, widgets for "top and popular comments", and the ability to "vote on comments without registering an account" (DB-2007-08), thereby lowering the threshold for anonymous participation. Features such as the "follow" and "vote" functi-ons were introduced within a context of trustworthy, creative, and commu-nitarian users. These users would reward each other for writing valuable comments by "voting", "rating", or "liking" them (DB-2007-10; DB-2009-03). From the start, these evaluative metrics were part of Disqus' backend reputation system called Clout DB-2007-11). Disqus describes Clout as its "numerical measure of a person's reputation across blogs in the Disqus di-scussion network. This number aids our spam and troll filtering. Your clout is determined by a few different things, but mainly by the popularity of the comments you make" (DFAQ-2007). Thus, the calculated reputation of a commenter directly informed the moderation process.

In 2009 "likes" gave "the commenter a positive reputation point on Disqus" to help "other readers discover great comments", and were mainly described as a discovery mechanism for quality content (DB-2009-03). In 2012, when a new major version of Disqus was introduced, "likes" were downgraded to "a lightweight virtual wink" and "votes" were reintroduced as the central metric at the core of new "reputation-based algorithms" for surfacing "qua-lity conversations" in an automated way (DB-2012-06).

Another example shows how popularity metrics were also claimed to serve other purposes than filtering conversations. A post from March 2008 explains their role in the commenting section for driving the conversation: "Another way to add some social spice to your blog is through Disqus' widgets. You can now add widgets to your blog to display your recent com-ments, top commenters, and even popular discussions" (DB-2008-03a). The semantics of the "social spice" is noteworthy: users should be spurred and incited rather than filtered, tamed, tempered, or moderated. Thus, in the late 2000s, popularity markers served different but related functions: on the one hand, they should spice up the conversation; and, on the other hand, they should assist in tempering the conversation by filtering out trolls and surfacing quality conversations. However, this process of filtering should be manually conducted by users voting on and liking comments; Moderation should be crowdsourced.[44]

Further, the proclaimed functions of these popularity markers have chan-ged over time. After 2010, these features focused less on surfacing quality and more on assisting with moderation. In 2012, they were (re)introduced as part of a new moderation interface, showing the moderator whether a user had a high, average, or low reputation based on the number of liked, flagged, or banned comments (DB-2012-03). As such, they were meant as additional information that moderators could use to (manually) evaluate contributions with the objective of "improving your community" (DB-2012-03). That is, Disqus further promoted popularity markers as means of assis-ted moderation.

44 Cf. Halavais.

Four years later these metrics would become key elements of Disqus' new strategy based on automating moderation on the level of the commenter. In a key blog post from April 2016, Disqus describes how its general approach towards moderation had changed. Until then, the company said its main focus had been on comment-centred moderation by developing tools for "deleting and editing comments". However, these tools "would never be at the core of a larger moderation strategy" (DB-2016-04a) because they required the engagement of a moderator at the level of an individual comment. To address this, Disqus claimed to change its general strategy for developing moderation tools by moving from a comment-centred to a commenter-centred approach, thereby shifting moderation from the object to the person (DB-2016-04a). To achieve this, the company details how it "introduced things like user reputation, user profiles, and trusted & banned user lists to help publishers moderate commenter-level behaviour rather than getting bogged down by individual comments" (DB-2016-04a).

In other words, the proclaimed functions of the popularity markers changed from a means to crowdsource moderation to more technologically advanced tools assisting with moderation. Thereby, Disqus increasingly focused on how their features could support moderators in dealing with problematic participation in ways that cannot be easily crowdsourced and more generally, shows how dealing with problematic participation became a central issue for website owners. This finding aligns with our quantitative distant reading of the blog via keyword analysis: Moderation becomes more and more an issue after 2012 and dominates the blog posts from 2016 onwards (see Figure 1). Although references to moderation and problematised forms of participation such as trolling can be found throughout our sample, the features to address this problematised participation change over time.

As Stevenson has shown, both the problematisation of commenting as well as the suggested technological solutions through valuation practices, are not fundamentally new. Slashdot had already introduced a scoring system of commenters in 1998, as a reaction to "various abuses of the comments system".[45] However, Slashdot's scores were initially based on the ratings of a selected group of moderators, and not crowdsourced to a community, and the moderators' "cumulative comments score" came to stand in for a "user's reputation" on the website.[46] Moreover, Slashdot described its features for moderation assistance in different semantics. The rating, which should help to curb abuse, was addressed with a playfully-ironic vocabulary: It was first called "alignment", which was a reference to a role-playing game and then renamed into "karma".[47] As shown next, the semantics Disqus applies to describe this type of participation fundamentally differs from this and changes over time.

Semantic shifts

From its beginning in 2007, Disqus was concerned with "powering better discussion", but over time, the notion of how one improves discussions changed. In a blog post from 2007, right before the launch of the service, Daniel Ha (one of the founders of Disqus) announced that the company was working on a place for "smart discussion" (DB-2007-04). He foregrounded the value of comments, as drivers of online discussions and conversations, for being "full of insightful information". However, he contended that these discussions also need moderation to "filter out the spam, trolling comments,

45 Stevenson, p. 622.

46 Stevenson, p. 622.

47 Stevenson, p. 622.

petty arguments, rambling digressions, and only keep the relevant gold" (DB-2007-04). Two aspects regarding Disqus' ideas about comment moderation are noteworthy here: The notion of filtering out gold from something irrelevant and the fact that they have been dealing with practices of trolling early on.

Trolling

The way Disqus has dealt with trolling — or at least the way they claimed to — changed over time. From the beginning in 2007, filtering out "trolling comments" (DB-2007-04) and "anti-troll" tools (DB-2007-10) have been an integral part of Disqus' technology. As the company claims, "When we started Disqus, one of the promises we wanted to make was the ability to curb comment trolling" (DB-2007-11) and that "Disqus was built to address such trolling" (DB-2007-12). For the company "trolling and harassment undeniably exists" and just needs "a better solution" (DB-2008-03b). Trolling is not necessarily considered harmful but instead is seen as an inherent aspect of online culture. It is positioned as a solvable issue by giving moderators tools for filtering out trolls and trolling comments.

One of these new tools to help moderators "more effectively determine trolls" was the "verified profile" (DB-2008-03c), which turned to user registration and user reputation to facilitate in the (manual) moderation process. In addition, moderators could search through comments "for common words used by trolls/spammers to moderate all their comments in one go" as well as whitelist and blacklist users (DB-2009-09).

In 2012, with the launch of a new major version, the community was more involved in handling "trollish comments" through voting (DB-2012-05). The company introduced a new voting system which "was meant to encourage richer discussions by letting communities curate the best comments" (DB-2012-05). However, as downvoting filtered out "trollish comments" Disqus positioned (up)voting as "a self-moderation tool for the community", suggesting that the community can solve trolling by itself through community moderation (DB-2012-05). This shows once more how Disqus promoted user valuations as a practice of crowdsourced moderation assistance. In addition, it also shows how trolls remained as an inherent part of the comment sections.

Until 2016, the blog posts mainly addressed publishers and moderators with tools for moderating trolls. This changed when users were given the possibility, next to "flagging a comment or reporting someone to a moderator", which requires further action from a moderator, to block someone directly to deal with harassment or trolling (DB-2016-06). In 2017, Disqus introduced two more tools for moderators "to fight trolls" and "improve discussion quality" with "shadow banning" and "timeouts", both of which are mechanisms to temper the conversation (DB-2017-05). In that year the semantics of disturbance shifted from "trolling" towards "toxic", and trolling (defined as "the main goal of the comment is to garner a negative response") became one of the four properties of "toxic content" (DB-2017-08).

Phillips and Milner show how in the late 2000s and early 2010s trolling was not only considered a normal and legitimate phenomenon by participants and observers of "Internet culture",[48] but also how it referred to a wide ran-

48 Cf. Barnes.

ge of Internet practices, including the production of memes. They criticise how this broad interpretation of the concept of trolling by various actors of Internet culture, including tech companies, journalists, and media scholars, has contributed to the normalisation of trolling and the valorisation of harmful participation practices.[49]

49 Phillips and Milner, pp. 53–59.

In our sample of Disqus blog posts, as well as in Barnes'[50] and Phillip's and Milner's[51] accounts, trolling became more problematised during the 2010s and was considered a harmful practice. This explains why in the post by Daniel Ha, published before Disqus' launch, trolling is mentioned as one of many kinds of practices producing more or less harmless waste, from which "gold" has to be filtered out. This implies a very specific model of separating value from waste; comments are depicted as gold nuggets that must be sifted out from masses of unwanted matter. However, these unwanted objects are in no way harmful to the wanted gold. Rather, the metaphor of gold panning already implies a factual necessity to explore bulks of dirt in a rough environment. This idea is reflected in the features that are implemented in the commenting system, with functions for "ratings and sorting" to "sort through the mess to get to the good stuff" (DB-2007-10). Initially then, trolling was considered harmless and, more than that, normalised as something one needs to cope with when digging for gold.

50 Barnes.
51 Phillips and Milner.

The semantics of comments as a form of toxic communication that must be combatted, which was addressed almost a decade later, implies a quite different model of differentiating value and waste. Two concepts of improving discussions alternate there: The earlier thought model, from the late 2000s, is concerned with exploring and filtering innocuous waste from valuable discoveries while the later model, from the late 2010s, centres around notions of conservation, saving the valued, evaluated, and whitelisted users and general public from contamination with toxic discussions. The challenge of improving quality was already central to the new Disqus version in 2012. However, this was not described with semantics that point to an urgency requiring action as with the diagnosis of "network climate change".[52]

52 Phillips and Milner, p. 19.

Health

Years before the first mention of toxicity on the Disqus blog, there was a shift towards another biologistic thought model, with the concept of the "healthy debate". This idea is first mentioned in January 2014, when a commenter is featured as someone "measured and thoughtful, but doesn't mind a little healthy debate" (DB-2014-01). The concept of healthiness employed here is associated with robustness, like a healthy immune system that can cope with viruses and other pathogens. In that sense, a healthy debate is one that can handle disruptive communication such as strong polemics. The implicit concept of healthiness is still in line with the thought model of gold mining: disruption, waste, and aggression are normal, and one must not be afraid of getting into contact with such phenomena of "Internet culture".

In an interview from August 2014, a different concept of healthiness is employed when a blogger is asked: "How do you encourage healthy debate among your readers?" (DB-2014-08). She answers that her respectful tone as a blog owner motivates users to be respectful, too: "My readers and comments like that the community is like that and it feels safe for them to comment" (DB-2014-08). Here, healthiness is not associated with robust-

ness but with safety. Instead of defining the healthy debate as being robust enough to cope with contamination, it is now defined by its protection from disrespectful communication. Whereas the former valorises a certain type of commenting practices, the latter pathologises them.

This concept of healthiness is further used in a post from April 2016, which was written after the Tay bot experiment, where Microsoft built a Twitter bot that automatically learned from the language used by its community, with the result that the bot mainly learned swearing words and other kinds of aggressive language (DB-2016-04b; Lee, 2016). Disqus' blog post asked, "what if Microsoft's Tay bot was unleashed in your community?" and discussed "best practices for growing a healthy community" (DB-2016-04b). That means that, also here, the healthy community is one that uses friendly language and the sick community is the one that trained the Tay bot. There are several other posts with the same concept of a healthy community.

In an internal hackathon in June 2016 Disqus' developers presented an idea for a tool called "Community Health Score" (DB-2016-07). This function would display the health score of a community based on three categories: activity, abuse, and (author/moderator) participation which "shape the health of any online community" (DB-2016-07). It demonstrated how internal developers drew on the biologistic notion of a community that can be more or less healthy or sick, and whose healthiness can, or even should, be monitored and diagnosed by its community managers, eventually requiring intervention. These notions of healthy communities, debates, and discussions remain dominant after 2014 and are especially present from 2016 to 2018. Afterwards, health refers mainly to physical health topics, such as food blogs with healthy vegan recipes.

Toxicity

The semantic field opened up by the use of metaphors from biology appears especially relevant for a history of online commenting. Metaphors have specific moral and political implications,[53] which Marianne van den Boomen studied extensively in the context of digital technologies.[54] She explains how metaphors for describing online sociality have a specific agency, as explained through the idea of the "online community". This notion "binds software, interface, page-space, and a group of users into a specific configuration that is both metaphorical and material: it orchestrates social conduct and assigns subject positions in the social ranking order that emerges from ongoing collective debate".[55] We outline these specific configurations and implications that are invoked by the company Disqus over time by using the metaphor of toxicity.

The word "toxic" first appeared in a blog post from April 2016 discussing whether comments "really add value" or are "just a necessary evil" and the related dilemma whether to "include a comments section and risk it being overtaken by Internet trolls" or not include such a section and lose "meaningful exchanges and valuable reader engagement" (DB-2016-04a). They outline a "scientific approach" taken by the new competing commenting system The Coral Project for analysing and solving problems related to online moderation (DB-2016-04a). Coral's approach applied "an academic rigour to studying topics like toxic conversations online, identifying trolls, and determining patterns of quality conversations" (DB-2016-04a). This

53 George Lakoff and Mark Johnson, *Metaphors We Live By*, ed. by With a new Afterword (Chicago, IL: University of Chicago Press, 1980) <https://press.uchicago.edu/ucp/books/book/chicago/M/bo3637992.html> [accessed 23 May 2022].

54 Marianne Van den Boomen, *Transcoding the Digital. How Metaphors Matter in New Media.* (Amsterdam: Institute of Network Cultures, 2014).

55 Van den Boomen, p. 162.

blog post then linked to an article written by media and communications scholar Nick Diakopoulos on The Coral Project website listing five academic articles under the headline: "How to identify 'toxic conversation'".[56] However, none of the articles explicitly mentioned "toxic" conversation or communication. This implies that Disqus either adopted this term from the academic community via Diakopoulos, from competitor Coral, or possibly signals a larger semantic shift of blogs and newspapers considering the web and its contents in transformation. One month later, another blog post spoke about how to keep a community "healthy" and mentions a "toxic user" for the first time (instead of toxic communication) (DB-2016-05), mirroring Disqus' general shift from a comment-centred to a commenter-centred approach to moderation.

In early 2017, the company was publicly accused of hosting extremist communities of "alt-right and white nationalists trolls" and of facilitating "hate speech" on websites such as Breitbart and Infowars which used the Disqus plugin.[57] This led to a petition signed by over 100,000 Internet users to force the company to deal with the issue.[58] In their response, Disqus acknowledged that hate speech is "the antithesis of community" and committed to better enforcing their terms of service, by investing in "new features for publishers and readers to better manage hate speech", and by emphasising how their new moderation panel helps to "identify and moderate comments based on user reputation" (DB-2017-02). While striving to facilitate "intelligent discussion", the company actively positioned itself as a comrade in the battle by "building tools for readers and publishers to combat hate speech" (DB-2017-02).

In April 2017, the company further reflected on their "commitment to fighting hate speech" and now referred to hate speech as "hateful, toxic content", thereby mentioning toxicity for a third time. The blog post outlines the company's "First Steps to Curbing Toxicity" and states: "While we do periodically remove toxic communities that consistently violate our Terms and Policies, we know that this alone is not a solution to toxicity" (DB-2017-04). This can be considered a further escalation, as toxicity is no longer applied to individual users but to communities as a whole.

The issue at stake for Disqus as a producer of commenting software is no longer one of exploring, discovering, surfacing, and highlighting valuable comment-nuggets among trolls, spam, and other kinds of allegedly harmless waste. Rather, comment sections have — according to their metaphorical description — seemingly turned into threatened and threatening environments. To counter these threats, the post introduces new and upcoming features such as "toxic content detection through machine learning" and new tools for advertisers who "do not want their content to display next to toxic comments" (DB-2017-04). As such, the toxic is not only considered a threat to the healthy debate but also to the monetizable comment space.

Disqus further elaborates on its concept of toxicity in August 2017. The company draws on earlier blog posts describing its attempts to "fight" hate speech and toxicity to explain its focus on "developing technology to facilitate good content" (DB-2017-08). It describes how the "most difficult and nebulous task" is to define what constitutes "'toxic' comments".

56 Coral, 'Artificial Moderation: A Reading List', *Coral by Vox Media*, 2016 <https://coralproject.net/blog/artificial-moderation-a-reading-list/> [accessed 23 October 2021].

57 Warzel.
58 https://actions.sumofus.org/a/disqus-you-host-breitbart-s-cesspool-of-a-comment-section-adopt-a-hate-speech-policy-now

According to the company, a comment is toxic if it meets at least two out of the four following properties: "abuse", "trolling", "lack of contribution", and the "reasonable reader property" (DB-2017-08).

In the rest of the post Disqus introduces a new technology that should help moderators cope with such toxic content, a so-called "toxicity barometer". The internal product name for this "toxicity-fighting technology" is "Apothecary". (DB-2017-08), which is "a combination of third-party services and our own homegrown data tools". The Apothecary tool signals Disqus' pivot to using machine learning techniques for identifying toxic content and its reliance on third-party technologies such as the Perspective API for doing so, which is further discussed below (DB-2017-08). This technology "spits out a toxicity score" that will be incorporated into the moderation panel and shown to other users. As such, the score serves a dual purpose: moderators and users can use it to decide whether they want to intervene or participate in the conversation, respectively.

The notion of the apothecary further expands on the implicit pathologisation of commenting: It promises to bring back "health" to communities by providing antidotes to the poison of contaminated speech. But the apothecary is also a suitable metaphor from Disqus' point of view, as they describe the main problem with providing technologies for semi-automated moderation. They explain how measures for dealing with unwanted commenting practices can vary strongly between different communities of practice and cannot be solved with a single solution: "Our publishers are a very diverse group so a 'one size fits all' approach to improving moderation simply wouldn't work" (DB-2017-08). The metaphor of the apothecary is then used to refer to personalised solutions of different patients — produced by and in the apothecary (in this case Disqus' Apothecary tool).

The semantics of toxic communication became even more prominent after the 2017 "Toxic Comment Classification Challenge". This competition was organised by the Conversation AI team of Google and Jigsaw, a unit within Alphabet aiming "to help improve online conversation" with artificial intelligence.[59] Their technologies include the Perspective API which uses "machine learning to identify 'toxic' comments" and which attributes comments with a "toxicity" score.[60] This technology became integrated into various commenting technologies and platforms, including Disqus, who now relies on this third-party artificial intelligence for detecting and managing toxic comments.[61]

One upside of the notion of the toxic appears to be that it can be measured in scores or barometers. Another, perhaps more central, specificity of this metaphor becomes evident when it is compared to the notion of healthy communities. As Keulartz and van der Weele have pointed out, "the health metaphor is akin to [...] practices such as rehabilitation—the treatment of severely diseased or disabled people with the purpose of re-socialization and reintegration into community life".[62] The toxic implies no such notion of reintegration. What is toxic has to be contained, preventing it from poisoning even more areas. Both metaphors pathologise and imply an urgency to act, but they suggest a different diagnosis and response. The "Apothecary" product promised to have a technical solution to "heal" the toxic — a promise that might have proven too optimistic.

59 https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

60 https://perspectiveapi.com/

61 Bernhard Rieder and Yarden Skop, 'The Fabrics of Machine Moderation: Studying the Technical, Normative, and Organizational Structure of Perspective API', *Big Data & Society,* 8.2 (2021).

62 Jozef Keulartz and Cor van der Weele, 'Framing and Reframing in Invasion Biology', *Configurations,* 16.1 (2008), p. 113.

However, not all metaphors used by Disqus for coping with unwanted practices imply pathologisation. When explaining how to deal with problematic commenters we also find that metaphors of pedagogy are employed. In a post from April 2016 about integrating community guidelines on your website, Disqus explained how to educate one's users: "Thus, you should find frequent opportunities to encourage readers to learn more about the rules for commenting, expectations and procedures for how moderation is handled, and how they can become a valued member" (DB-2016-04c). In May 2017, Disqus introduced the "shadow banning" and "timeouts" functions which give the banned user "a chance to cool off" and "an opportunity to improve their behavior and return to the community" (DB-2017-05). By doing so, Disqus uses the language of disciplining maladjusted children and imports classic pedagogical techniques for behavioural management using time-out interventions.

Thus, next to metaphors pathologising certain types of user participation, Disqus has drawn upon several semantic fields. Already the notion of "combating hate speech" opens up a field of its own, that of warfare, defining an enemy that urgently needs to be battled. Apparently, the semantic transformations mirror the evolving views on the state of commenting and the role of Disqus in shaping commenting and its moderation, testing and trying to find solutions for "powering better discussions". However, the most prominent semantic field was of toxicity, which implicitly accepts that there is no solution to be found for certain commenting practices, except for the defensive strategy of preventing them to spill over to even more users and even more websites.

Conclusion: Increasing Platform Interventionism and Metaphors of Urgency

Over the past 15 years, Disqus has changed its features for managing and moderating commenting as well as its language on commenting in several respects. On a general level, the company has become less concerned with facilitating and intensifying discussions, whilst crowdsourcing issues with problematic participation to a trusted community of users, and more with helping moderators to cope with these issues or intervene in them. Over time, we find that the practices and technologies of moderation have become more differentiated with the introduction of new policies, features, and the integration of third-party machine-learning technologies such as the Perspective API which assist in the process of moderation by semi-automating interventions.

Platforms and their moderation practices have always been interventive while presenting themselves as neutral.[63] However, this article has shown how Disqus, as one of the most important commenting platforms, has always been explicitly interventive. It has provided website owners with commenting and moderation functions to become increasingly interventive over time. This poses questions about the specificity of commenting platforms. Have they always been different from other platforms in terms of their "interventiveness"? And how does their increasing interventionism serve as a media-historical predecessor for other platforms, which are also becoming more and more openly interventive (but only years after the commenting platforms)?

63 Cf. Gillespie; Van Dijck.

Next to increasing modes of platform interventionism, we observed semantic shifts in how this interventionism is framed. Based on our research, we cannot say in how far these shifts in semantics can be considered effects of changes in the platform's policies or politics, or whether they are driven by "larger" political or historical processes (or have contributed to those processes). However, comment sections have increasingly become places of conflict. This development has been countered with more moderation, automated assistance, and other measures to temper commenting practices.

Even if we cannot make out direct causes and effects of the semantic shifts, we can determine what they do. The employed metaphors from diverse semantic fields, ranging from medicine to pedagogy to warfare, serve to legitimise the platform's interventions as urgent because the situation requires it. They imply that one has to act immediately, no matter what earlier hopes and beliefs about the future of participatory digital culture were. Trolling and other disruptive practices have long been considered harmless, or at least controllable parts of digital culture, but at a certain point Disqus started to describe them with metaphors of urgency. These metaphors imply a legitimation to suspend fundamental convictions or to radically change them. Furthermore, the integration of the Perspective API into Disqus' technology, demonstrates how problematised practices and their naturalistic discourse have become inscribed into commenting software.

Our sample represents more than a solitary case in this respect. Other competing commenting systems such as OpenWeb even go so far as to state that there is "a crisis of toxicity online" and that the aim of their company is "building an open, healthier web".[64] In the field of Media Studies such semantics of contamination, pollution, and toxicity are also increasingly used to problematise contemporary histories of digital cultures.[65] Further research into web histories and earlier web technologies could help to better understand how and why comment sections specifically and the web in general have turned into spaces of urgent action over the last decade.

Acknowledgements

The authors would like to thank Yarden Skop and Lisa Gerzen for their helpful feedback on earlier versions of this paper and Martina Schories for her assistance with the graphics.

Declaration of interest

The authors report there are no competing interests to declare.

Bibliographical Note

Johannes Paßmann is Junior Professor of History and Theory of Social Media and Platforms at Ruhr University Bochum and Principal Investigator at University of Siegen's SFB 1472 Transformations of the Popular. Previously, he was Post-Doc at the Digital Media & Methods team at Siegen University. His research focuses on theory, history, and aesthetics of social media.

64 https://www.openweb.com/

65 Phillips and Milner.

Anne Helmond is Assistant Professor of New Media and Digital Culture at the University of Amsterdam and Research Associate at SFB 1472. Her research interests include the history and infrastructure of platforms and apps, digital methods, platformization, and web history.

Robert Jansma is a research associate and doctoral candidate at the University of Siegen in the SFB 1472. He is from a computer science background, specialised in software engineering for preserving and studying digital heritage. His research interests include commenting systems, web and software histories, and archiving methods.

References

Al Maruf et al., Hasan, 'Human Behaviour in Different Social Medias: A Case Study of Twitter and Disqus' (presented at the IEEE/ACM International Conference 2015, Paris, France, 2015) <http://dx.doi.org/10.1145/2808797.2809395>

Barnes, Renee, *Uncovering Online Commenting Culture: Trolls, Fanboys and Lurkers* (Palgrave Macmillan, 2018) <https://doi.org/10.1007/978-3-319-70235-3>

Ben-David, Anat, and Oren Soffer, 'User Comments across Platforms and Journalistic Genres', *Information, Communication & Society*, 22.12 (2019), 1810–29 <https://doi.org/10.1080/1369118X.2018.1468919>

Blood, Rebecca, 'How Blogging Software Reshapes the Online Community', *Communications of the ACM*, 47.12 (2004), 53–55

Bruns, Axel, *Gatewatching: Collaborative Online News Production* (New York, NY: Peter Lang, 2005)

BuiltWith, 'Comment System Usage Distribution in the Top 1 Million Sites', 2021 <https://trends.builtwith.com/widgets/comment-system> [accessed 9 September 2021]

Burgess, Jean, 'From "Broadcast Yourself" to "Follow Your Interests": Making over Social Media', *International Journal of Cultural Studies*, 18.3 (2015), 281–85 <https://doi.org/10.1177/1367877913513684>

Charmaz, Kathy, *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis* (Thousand Oaks, CA: SAGE, 2006)

Collins, Luke, and Brigitte Nerlich, 'Examining User Comments for Deliberative Democracy: A Corpus-Driven Analysis of the Climate Change Debate Online', *Environmental Communication*, 9.2 (2015), 189–207 <https://doi.org/10.1080/17524032.2014.981560>

Coral, 'Artificial Moderation: A Reading List', *Coral by Vox Media*, 2016 <https://coralproject.net/blog/artificial-moderation-a-reading-list/> [accessed 23 October 2021]

Diakopoulos, Nicholas, *Automating the News: How Algorithms Are Rewriting the Media, Automating the News* (Harvard University Press, 2019) <https://doi.org/10.4159/9780674239302>

Disqus, 'About Disqus', 2021 <https://disqus.com/company/> [accessed 9 September 2021]

Frischlich, Lena, Svenja Boberg, and Thorsten Quandt, 'Comment Sections as Targets of Dark Participation? Journalists' Evaluation and Moderation of Deviant User Comments', *Journalism Studies*, 20.14 (2019), 2014–33 <https://doi.org/10.1080/1461670X.2018.1556320>

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

Glaser, Barney G., and Anselm Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research* (New York: Aldine, 1967)

Goodman, Emma, 'Online Comment Moderation: Emerging Best Practices' (World Association of Newspapers, 2013) <https://netino.fr/wp-content/uploads/2013/10/WAN-IFRA_Online_Commenting.pdf>

Halavais, Alexander, 'Do Dugg Diggers Digg Diligently?', *Information, Communication & Society*, 12.3 (2009), 444–59 <https://doi.org/10.1080/13691180802660636>

Hoffmann, Anna Lauren, Nicholas Proferes, and Michael Zimmer, '"Making the World More Open and Connected": Mark Zuckerberg and the Discursive Construction of Facebook and Its Users', *New Media & Society*, 20.1 (2018), 199–218 <https://doi.org/10.1177/1461444816660784>

Keulartz, Jozef, and Cor van der Weele, 'Framing and Reframing in Invasion Biology', *Configurations, 16.1 (2008)*

Kim, Jin Woo, Andrew Guess, Brendan Nyhan, and Jason Reifler, 'The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity', *Journal of Communication*, jqab034, 2021 <https://doi.org/10.1093/joc/jqab034>

Ksiazek, Thomas B., and Nina Springer, *User Comments and Moderation in Digital Journalism: Disruptive Engagement* (New York: Routledge, 2020) <https://doi.org/10.4324/9780429276132>

Lakoff, George, and Mark Johnson, *Metaphors We Live By*, ed. by With a new Afterword (Chicago, IL: University of Chicago Press, 1980) <https://press.uchicago.edu/ucp/books/book/chicago/M/bo3637992.html> [accessed 23 May 2022]

Nelson, Laura K., 'Computational Grounded Theory: A Methodological Framework', *Sociological Methods & Research*, 49.1 (2020), 3–42 <https://doi.org/10.1177/0049124117729703>

Omernick, Eli, and Sara Owsley Sood, 'The Impact of Anonymity in Online Communities', in *2013 International Conference on Social Computing,* 2013, pp. 526–35 <https://doi.org/10.1109/SocialCom.2013.80>

Paßmann, Johannes, 'Medien-theoretisches Sampling. Digital Methods als Teil qualitativer Methoden', *Zeitschrift für Medienwissenschaft*, 13.2 (2021), 128–40 <https://doi.org/10.25969/MEDIAREP/16779>

Phillips, Whitney, and Ryan M. Milner, *You Are Here: A Field Guide for Navigating Polarized Speech, Conspiracy Theories, and Our Polluted Media Landscape* (Cambridge, Massachusetts ; London, England: The MIT Press, 2021)

Reagle, Joseph M., *Reading the Comments: Likers, Haters, and Manipulators at the Bottom of the Web* (MIT Press, 2016)

Rieder, Bernhard, and Yarden Skop, 'The Fabrics of Machine Moderation: Studying the Technical, Normative, and Organizational Structure of Perspective API', *Big Data & Society*, 8.2 (2021) <https://doi.org/10.1177/20539517211046181>

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

Rogers, Richard, 'Debanalizing Twitter: The Transformation of an Object of Study', in *Proceedings of the 5th Annual ACM Web Science Conference* (ACM, 2013), pp. 356–65 <https://doi.org/10.1145/2464464.2464511>

Rowe, Ian, 'Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms', *Journal of Broadcasting & Electronic Media*, 59.4 (2015), 539–55 <https://doi.org/10.1080/08838151.2015.1093482>

Seifert, Josef W., 'Moderieren', in *An-Gewandte Sozialpsychologie*: Das Praxishandbuch, ed. by A. E. Auhagen and H. W. Bierhoff (Weinheim: Beltz PVU, 2003), pp. 75–87

Shin et al., Sungho, 'Analytics on Online Discussion and Commenting Services', in *Human Interface and the Management of Information. Information and Interaction Design.* (Berlin, Heidelberg: Springer, 2013), pp. 250–58 <https://link.springer.com/book/10.1007%2F978-3-030-39469-1>

Stevenson, Michael, 'Slashdot, Open News and Informated Media: Exploring the Intersection of Imagined Futures and Web Publishing Technology', in *New Media, Old Media*: A History and Theory Reader, ed. by Wendy Hui Kyong Chun, Anna Watkins Fisher, and Thomas Keenan (London: Routledge, 2016), pp. 616–30

Swales, John M., and Priscilla S. Rogers, 'Discourse and the Projection of Corporate Culture: The Mission Statement', *Discourse & Society*, 6.2 (1995), 223–42 <https://doi.org/10.1177/0957926595006002005>

Trischler, Ronja, and Beatrice Perera, 'Digitale Rituale Moderieren' (unpublished Paper presented at the Workshop Digitale Rituale, Frankfurt a.M.)

Van den Boomen, Marianne, *Transcoding the Digital. How Metaphors Matter in New Media.* (Amsterdam: Institute of Network Cultures, 2014)

Van Dijck, José, *The Culture of Connectivity: A Critical History of Social Media* (New York: Oxford University Press, 2013)

Warzel, Charlie, 'How The Alt-Right Manipulates The Internet's Biggest Commenting Platform', BuzzFeed, 2018 <https://www.buzzfeednews.com/article/charliewarzel/how-the-alt-right-manipulates-disqus-comment-threads> [accessed 14 September 2021]

Weech, Shelton, 'Analyzing Organizational Values Over Time: A Methodological Case Study', in *Proceedings of the 38th ACM International Conference on Design of Communication*, SIGDOC '20 (New York, NY, USA: Association for Computing Machinery, 2020), pp. 1–6 <https://doi.org/10.1145/3380851.3416760>

Wolfgang, J David, 'Taming the "Trolls": How Journalists Negotiate the Boundaries of Journalism and Online Comments', *Journalism*, 22.1 (2021), 139–56 <https://doi.org/10.1177/1464884918762362>

de Zeeuw, Daniel, Sal Hagen, Stijn Peeters, and Emilija Jokubauskaite, 'Tra-

cing Normiefication: A Cross-Platform Analysis of the QAnon Conspiracy Theory', *First Monday*, 2020 <https://doi.org/1603910724>

Ziegler, Albert, 'Wer Moderieren Will, Muß Maß Nehmen Und Maß Geben.', *Gruppendynamik-Zeitschrift Für Angewandte Sozialpsychologie*, 23.3 (1992), 215–36

Appendix

In-text citation	URL
DB-2007-04	https://web.archive.org/web/20101015210512/http://blog.disqus.net/2007/04/10/bhl-presents/
DB-2007-08	https://web.archive.org/web/20090604212722/http:/blog.disqus.net/2007/08/28/beta-opens-up-some-more-plus-new-changes/
DB-2007-10	https://web.archive.org/web/20100213051448/http://blog.disqus.net/2007/10/30/disqus-launches-in-public-beta/
DB-2007-11	https://web.archive.org/web/20100409101246/http://blog.disqus.net/2007/11/20/tickets-tracking-trolling-and-turkey/
DB-2007-12	https://web.archive.org/web/20100606013706/http://blog.disqus.net/2007/12/07/on-content-ownership-or-how-disqus-received-a-lawsuit-threat/
DFAQ-2007	https://web.archive.org/web/20071226063520/http://disqus.com/faq/
DB-2008-03a	https://web.archive.org/web/20100220192444/http:/blog.disqus.net/2008/03/14/disqus-releases-beta-2-%e2%80%93-new-features-lots-of-bug-fixes-improved-performance/
DB-2008-03b	https://web.archive.org/web/20100615060933/http://blog.disqus.net/2008/03/10/what-if-anonymous-comments-were-illegal/
DB-2008-03c	https://web.archive.org/web/20101111172122/http://blog.disqus.net/2008/03/25/statistics-on-verified-profiles/
DB-2009-03	https://web.archive.org/web/20100511091655/http://blog.disqus.net/2009/03/26/upgraded-interface-and-performance/
DB-2009-09	https://web.archive.org/web/20100516232100/http://blog.disqus.net/2009/09/01/guide-new-moderation/
DB-2012-03	https://web.archive.org/web/20120315211302/http://blog.disqus.com/post/18875154463/the-all-new-disqus-moderation-introducing-user
DB-2012-05	https://web.archive.org/web/20120526000359/http://blog.disqus.com/post/23634202841/d12-the-next-steps
DB-2012-06	https://web.archive.org/web/20120625100051/http://blog.disqus.com/post/25017922977/the-new-disqus-2012
DB-2014-01	https://blog.disqus.com/commenter-spotlight-hamish-lamont
DB-2014-08	https://blog.disqus.com/3-tips-for-sparking-rich-discussions-in-your-community-ft.-nikki-of-styling-you
DB-2016-04a	https://blog.disqus.com/are-you-doing-moderation-all-wrong
DB-2016-04b	https://blog.disqus.com/what-if-microsofts-tay-bot-was-unleashed-in-your-community
DB-2016-04c	https://blog.disqus.com/4-best-practices-for-integrating-community-guidelines-in-your-site
DB-2016-05	https://blog.disqus.com/when-does-moderation-become-censorship
DB-2016-06	https://blog.disqus.com/user-blocking-is-now-available-on-disqus

In-text citation	URL
DB-2016-07	https://blog.disqus.com/disqus-hacks-the-roadmap-june-2016-recap
DB-2017-02	https://blog.disqus.com/our-commitment-to-fighting-hate-speech
DB-2017-04	https://blog.disqus.com/first-steps-to-curbing-toxicity
DB-2017-05	https://blog.disqus.com/introducing-shadow-banning-and-timeouts
DB-2017-08	https://blog.disqus.com/developing-toxicity-fighting-technology-at-disqus
DB-2017-12	https://blog.disqus.com/disqus-and-zeta