The Block Log: 20 Years of Content Moderation on Wikipedia

Ryan McGrady, University of Massachusetts Amherst

Abstract

This study examines two decades of user blocking on the English Wikipedia to understand how a volunteer-run, non-profit platform has adapted its content moderation practices in response to increasing visibility amid declining participation. Analyzing more than 20 million block log entries from 2004 to 2024, the study identifies shifts in block frequency, duration, and stated rationales. A significant increase in preemptive, automated blocking of open proxies since 2020 accounts for most block activity, but excluding these reveals a broader trend toward longer blocks and vaguer rationales such as “disruption.” These patterns suggest that volunteers are scaling labor through automation and normative adjustment, trading openness for efficiency and stability. Wikipedia's blocking trends help to contextualize governance pressures on volunteer-run knowledge platforms.

Keywords: Wikipedia, content moderation

Introduction

Content moderation has become an increasingly prominent topic in public conversations about communication platforms, free speech, politics, and knowledge itself. A series of high-profile decisions by large technology companies have moved it from an obscure internal process to a focal point in the culture wars. But while researchers have thoroughly examined content moderation on platforms like Facebook and Twitter (Gillespie, 2018; Roberts, 2019; Gorwa et al., 2020), less attention has been paid to non-profit, volunteer-run systems.

Consistently ranking in the top ten most visited websites, Wikipedia plays a significant role in dissemination of knowledge and feeds a wide range of external information services like LLMs and digital assistants (Ford, 2022; McDowell and Vetter, 2022). Content moderation on Wikipedia is ripe for study not only because of its influence, but also because its mechanisms are radically transparent when compared to the opaque practices of commercial platforms (Gillespie, 2018). This study concerns the English Wikipedia, the oldest, largest, and most influential of the Wikimedia projects.

Moderation concerns the evaluation of user-generated content to ensure it is appropriate for a particular website. It presupposes a sometimes vast sociotechnical system of humans, policies, interfaces, and a wide range of software mechanisms to execute or assist in evaluation, decision-making, intervention, or approval. According to Tarleton Gillespie, moderation is central to what it means to be a platform. Studying for-profit platforms in particular, he refers to the rules governing moderation decisions as "at best, reasonable compromises — between users with different values and expectations, as well as between the demands of users and the demands of profit" (Gillespie, 2018, p. 12). In contrast, Wikipedia's rules and practices are not only public-facing, but the public is invited to participate in their enforcement. They are still compromises, but not between user expectations and profit; instead, the compromise is between content quality and core principles of open participation — the stricter the rules and norms for quality, the more barriers there are to the "anyone can edit" ethos.

Grimmelmann’s taxonomy (2015) offers a complementary perspective that helps to situate blocking in a broader moderation system. Concerned more with the practicalities of moderation than their relevance to business, law, technology, or politics, Grimmelmann defines moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" (p. 47) and distinguishes parts of an involved community (members, content, and infrastructure), the goals of moderation (productivity, openness, and keeping low costs), types of moderation actions (including exclusion), and structural characteristics of communities that influence effectiveness of techniques.

How Wikipedia undertakes content moderation is different from most of the high-profile platforms for several reasons. First and most obviously, Wikipedia is not for profit. Decisions about content on Wikipedia are guided not by investor interests but a commitment to producing a public resource. Second, while the Wikimedia Foundation (WMF) develops Wikipedia's core software and hosts its servers, it rarely intervenes in daily operations and does not own the content. Moderation decisions are made by volunteers, not staff. All users can edit article content, and a subset of trusted users called administrators are given advanced permissions — mainly, block, protect, or delete. For the first few months of Wikipedia's existence, founder Jimmy Wales handled these tasks personally, appointing the first group of administrators in October 2001 (Schiff, 2006). Since 2003 appointments have been made by a democratic process.

To understand how Wikipedia has adapted its content moderation practices in response to increasing visibility and declining participation, this study examines one administrator action in particular: blocking, the removal of a user's ability to edit pages. This leads to three research questions:
RQ1: How has blocking activity on the English Wikipedia changed from 2004 to 2024?
RQ2: How have the rationales provided by administrators evolved, and what do these changes reveal about shifting norms and priorities?
RQ3: How have block durations changed over time, and what do these changes suggest about the community's strategies for scaling volunteer labor?

Background

Over two decades, Wikipedia has evolved from a quirky encyclopedia project to a central global knowledge resource or even a bulwark against misinformation (McDowell and Vetter, 2020). At the same time, major companies like Google, Facebook, YouTube, Amazon, and OpenAI increasingly integrate Wikipedia content (Jankowski, 2023; Brown et al., 2020). Its heightened visibility has, in turn, made it an attractive target for the exercise of money and power: self-promotion, marketing firms, political advocacy, and even state-sponsored Wikipedia influence campaigns (Miller, et al., 2022). This dual popularity and apparent vulnerability has raised interest in the site's content moderation.

To understand why Wikipedia does not descend into a free-for-all of political agendas and advertising, researchers have emphasized its strong set of shared norms, adherence to a body of community-written rules, and an emphasis on socialization (Choi et al., 2010; McGrady, 2013; Grimmelmann, 2015; Morgan and Halfaker, 2018). Over time, Wikipedia contributors codified a wide range of best practices, eventually taxonomized by domain (content, conduct, style, procedure, etc.) and degree of flexibility ("policies" are core principles, "guidelines" are best practices, "essays" are supplemental or represent one or many perspectives). Additional layers of policy emerged not just to further explain or bridge policies, but also to ensure they were enforced faithfully (McGrady, 2009).

Rules have long been understood to play an essential role in Wikipedia's governance (e.g. Kriplean, et al., 2007; Butler, et al., 2008). Like its encyclopedia articles, Wikipedia's rules are open to continuous revision In Keegan & Fiesler's 2017 analysis of this rule-based governance over Wikipedia's first fifteen years, the authors distinguish the rules which document standards ("rules-in-form") from dominant practices ("rules-in-use"). They found revision of the rules-in-form persisted even as edits to policy pages slowed, shifting activity to be more deliberative. Surprisingly, a dynamic ruleset did not appear to disrupt the community's ability to moderate, which the authors attribute to the stability of rules-in-use.

But the functionality of the system is fundamentally dependent on a robust volunteer base. When its popularity exploded in 2005-2006, its community of volunteer editors grew, too, but not to the extent needed to contend with an influx of activity. Increasingly, volunteers spent a larger proportion of their time on "non-direct work" (Kittur, et al., 2007) to facilitate decentralized governance. Not only have pageviews remained consistent (FIGURE 1), but there is more content than ever to maintain (FIGURE 2). The number of editors peaked in 2007 and has decreased ever since (FIGURE 3) (Wikimedia Foundation, n.d.). Across different language Wikipedias, new user registrations have declined not just since 2007, but 35% just since 2019 (Shah-Quinn, 2025). The good news for Wikipedia is that the number of *active* users (those who make at least five edits in a month) has been relatively stable (FIGURE 4) and the overall number of user edits per month has been increasing since a low in 2013 (FIGURE 5) (Wikimedia Foundation, n.d.). Perhaps most consequential, however, is the decline in administrators (FIGURE 6). Administrators, the trusted users with access to blocking, deletion, and protection tools,[1] fell from nearly 1,800 in 2011 to 846 in March 2025.[2] There are a wide range of potential reasons for this reduction, ranging from shifting standards for promotion to burnout (Asikin-Garmager et al., 2025), but the result is fewer people to carry out core moderation tasks even as the workload grows.

[1] https://en.wikipedia.org/wiki/Wikipedia:Administrators

[2] https://en.wikipedia.org/wiki/User:NoSeptember/admincount#Historical_admin_count

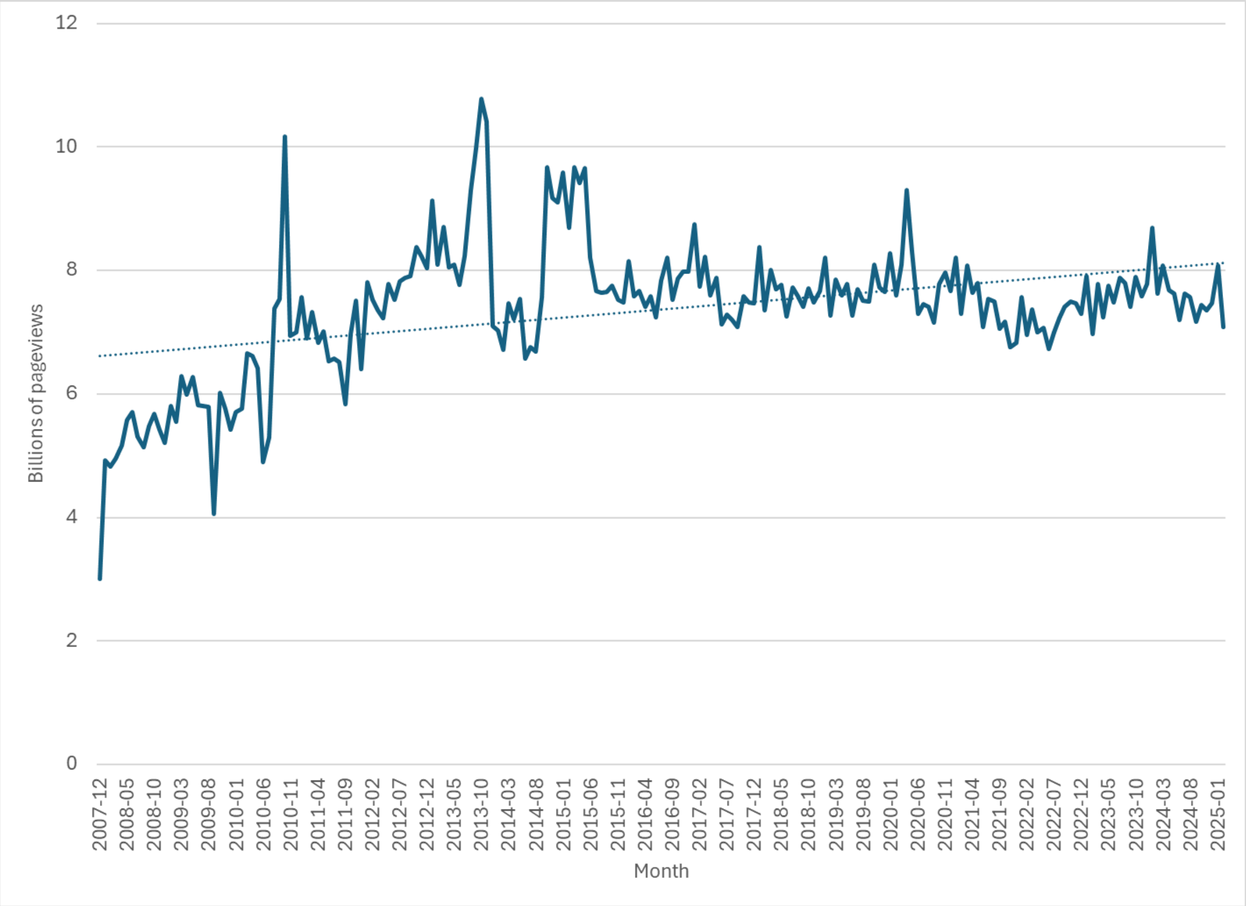

Figure 1: Monthly pageviews of the English Wikipedia[3]

[3] Data retrieved from https://pageviews.wmcloud.org/siteviews and https://stats.wikimedia.org

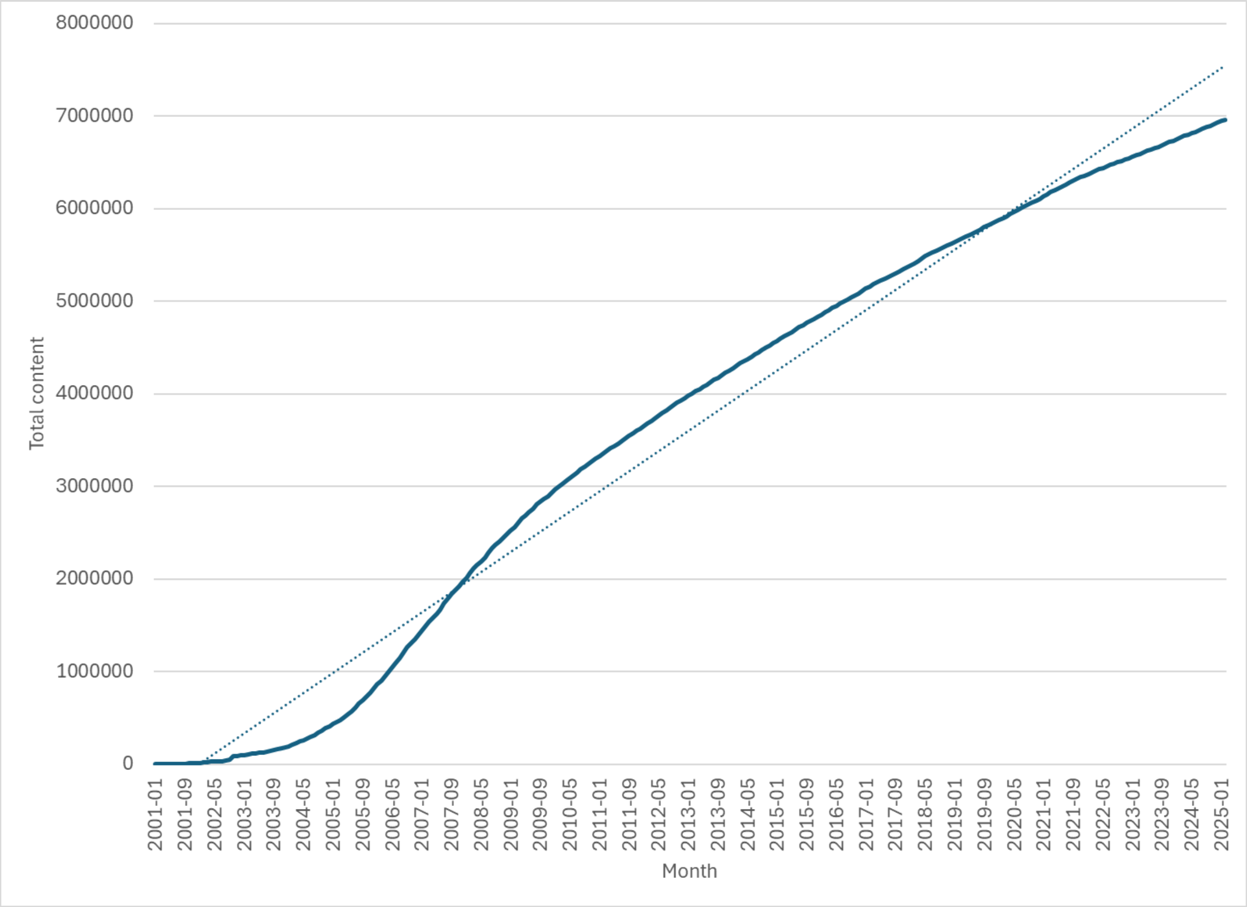

Figure 2: Total content pages

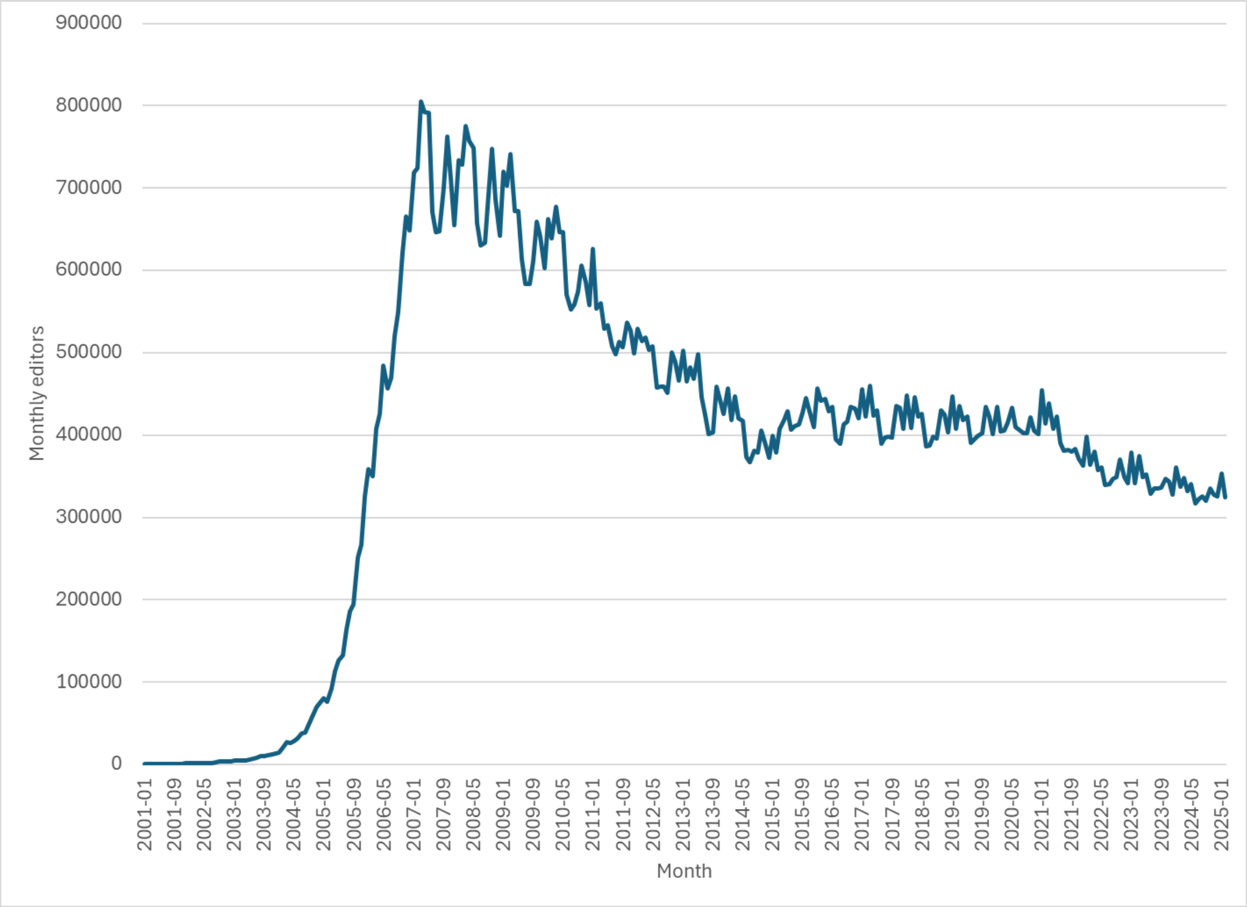

Figure 3: Monthly editors

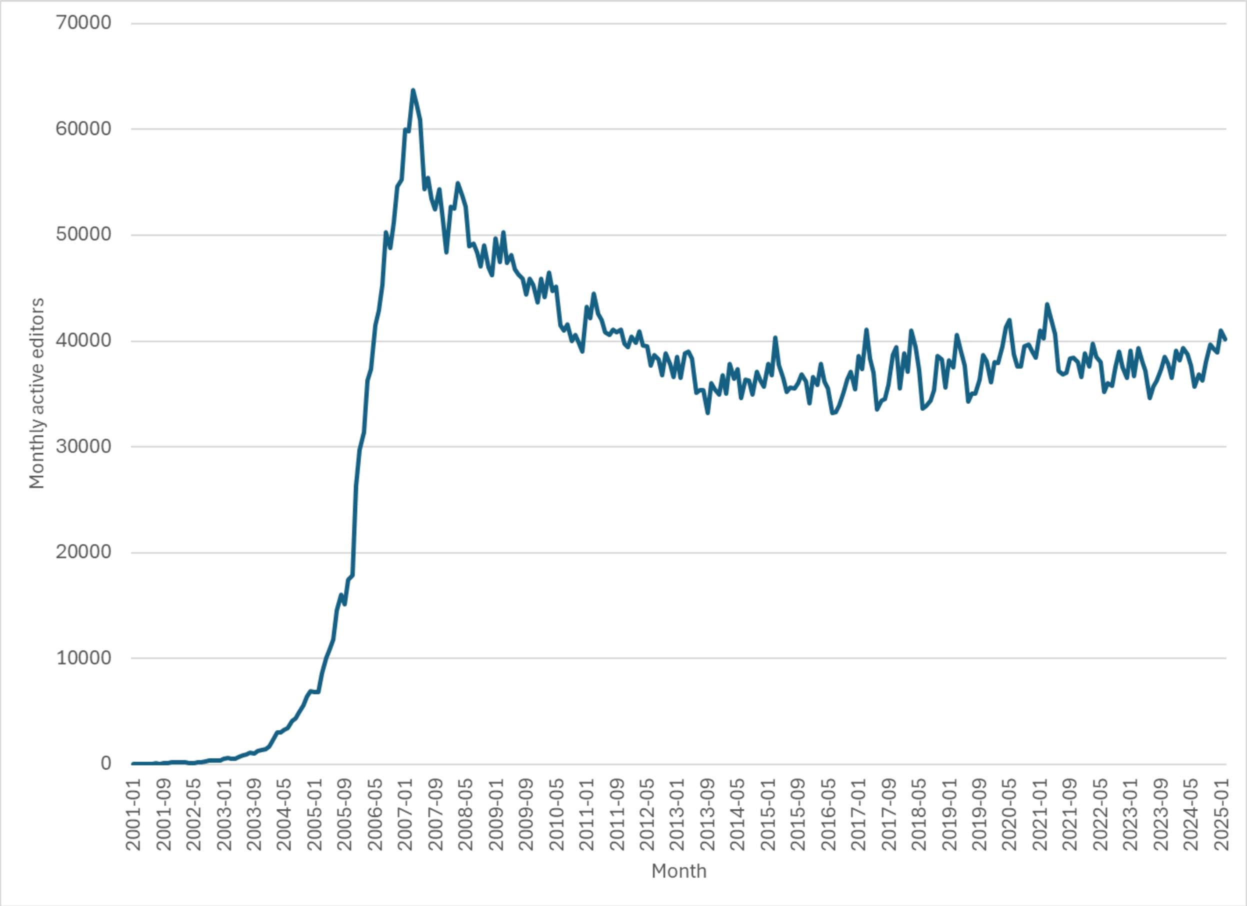

Figure 4: Active editors over time

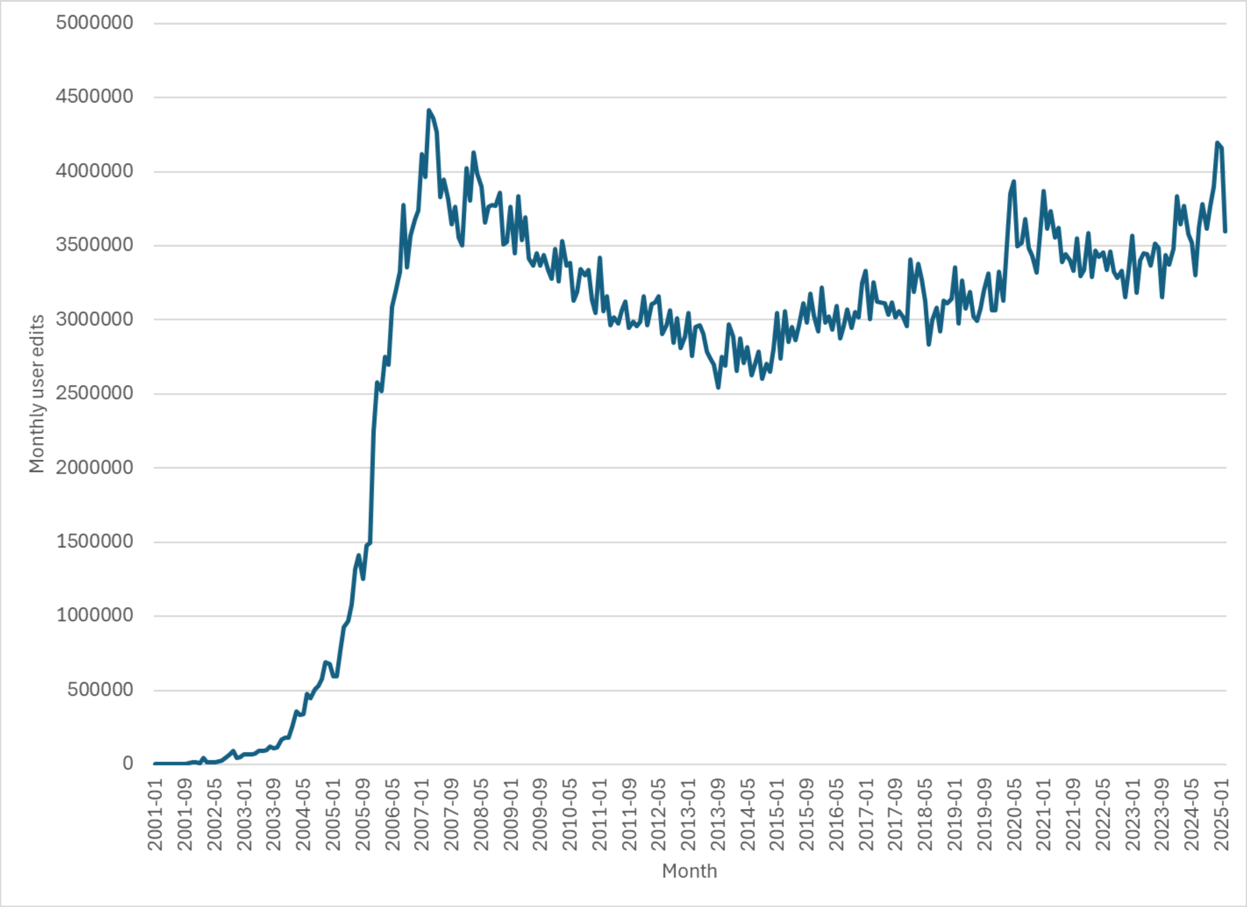

Figure 5: Monthly user edits

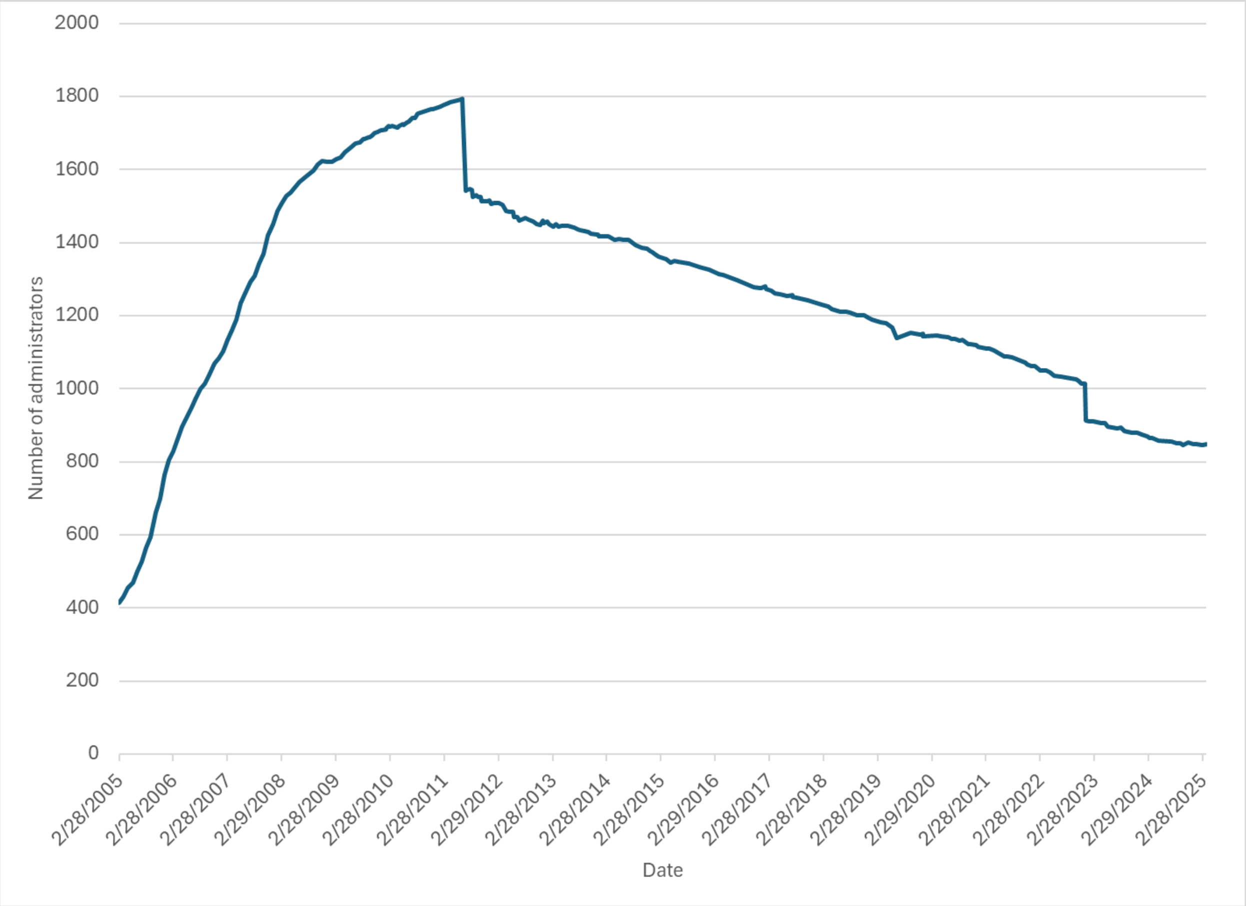

Figure 6: Number of administrators over time. The sharp drop in 2011 is due to administrator activity requirements being enacted, removing a large number of inactive administrators.

Unlike commercial platforms, Wikipedia cannot simply hire more moderators. It can try to recruit volunteers, which it does through on-wiki invitations and offline events, but recruiting is slower and less predictable than hiring, and the community has had to find other ways to bridge the gap between volunteers, activity, and popularity.

One approach to scale volunteer labor is through technical tools that automate or assist with common tasks. Tools like Huggle, STiki, RedWarn, and WikiLoop DoubleCheck flag obviously harmful “vandalism”[4] as well as potentially problematic edits in real time, sometimes gamifying the repetitive task of undoing bad edits.[5] Other tools facilitate page moves, article deletion, and communication.[6] Geiger published several studies about the role of bots in Wikipedia, arguing that understanding the work done by non-humans is essential to understanding Wikipedia governance (e.g. Geiger, 2011). His study of data from 2009 found bots (automated edits) had increased from 2-4% of edits in 2005-2006 to 16.33% of all edits, with another 12.16% of edits "semi-automated" (humans making decisions aided by software or scripts) (Geiger, 2009). In

[4] https://en.wikipedia.org/wiki/Wikipedia:Vandalism
[5] https://en.wikipedia.org/wiki/Wikipedia:Cleaning_up_vandalism
[6] https://en.wikipedia.org/wiki/Wikipedia:Tools

2024, the percentage of bot edits was up to 28.02%.[7] The influence of bots is probably underestimated, both in terms of affecting articles and the editing community (Jiang and Vetter, 2019). For administrator work, however, bots can only block users in very rare circumstances.[8] Tools can assist humans and make reports for humans to judge and act upon, but most administrator work is still done by humans (Asikin-Garmager et al., 2025). While effective, many of these technical interventions make the environment less welcoming to newcomers, with less tolerance for mistakes (Halfaker et al., 2012).

Another approach to scaling volunteer labor is by tweaking policy ("rules-in-form") or, commonly, by shifting normative applications of those policies ("rules-in-use") (Keegan & Fiesler, 2017). Administrators are commonly regarded as "janitors," merely enforcing community consensus, but they exercise a considerable amount of judgment and discretion in day-to-day operations (Jemielniak, 2013; Tkacz, 2015; Keegan & Fiesler, 2017). As with infrastructural changes, normative shifts often trade some openness for more productivity or lower costs (in the sense of volunteer time).

Together, these interventions represent styles of "boundary-work," a term from science studies describing the ways in which scientists continually struggle to redefine what counts as science based on practical needs (Gieryn, 1983). On Wikipedia, contributors implicitly or explicitly redefine what content or users should be included or excluded based on rules, norms, and practical demands of the present moment. While the effect of these changes is not always to create a stricter environment or to concentrate power (sometimes the opposite is true), even efforts to increase openness increase bureaucratization — or "self-organizing bureaucratization" (Rijshouwer et al., 2023) — that is increasingly difficult to reduce.

In this study, I focus on one administrator task in particular: blocking. On most platforms, "blocking" primarily serves an interpersonal filtering function, preventing one user from communicating with another. Blocklists serve an essential function countering harassment on Twitter, for example (Jhaver et al., 2018). In John's taxonomy of "disconnectivity" features, blocking on Wikipedia is closest to muting, preventing a user from effecting any visible communication (2024). But while blocking does play an important role in addressing harassment and toxic behavior on Wikipedia (Wulczyn, et al., 2017; Schluger, et al., 2022), it is not strictly, or even primarily, interpersonal. Blocking is the act of revoking or restricting the editing privileges of a user, moderating content by limiting who can create or modify it, or what Grimmelmann would call "exclusion." A block can be applied to a registered account or to an IP address and most commonly restrict edits to all pages except the user's own talk page, which is used for discussion and appeals. Following Ostrom's principle of graduated sanctions (1990; Chandrasekharan, et al., 2022), blocks can be for a short period or indefinite, and are intended to prevent harm rather than punish users for bad behavior, although the distinction is often blurry.[9] Traditionally a last resort following warnings, blocks can nonetheless be enacted more quickly than other corrective measures. In a context where standards are increasing but labor is

[7] https://en.wikiscan.org/

[8] https://en.wikipedia.org/wiki/Wikipedia:Bot_policy

[9] https://en.wikipedia.org/wiki/Wikipedia:Blocking_policy

decreasing, we might expect to see a larger number of blocks or more severe blocks to deal with bad actors (or potential bad actors) more quickly.

There is some existing research on blocking on Wikipedia. Chang & Danescu-Niculescu-Mizil studied the trajectory of users following a temporary block, acknowledging that blocking can be necessary for community functioning even while it can alienate potentially constructive contributors. They found that the characteristics of the blocked person and their perception of fairness (including the duration of the block) determined whether the user tended towards “redemption, recidivism, [or] departure” (2019). The issues of fairness and speech recur frequently in broader literature about content moderation.

On Wikipedia, nearly every action leaves a digital trace, including administrative actions, which are stored in publicly visible logs. This degree of transparency sets it apart from the opaque moderation practices of other platforms. For this study, I retrieved twenty years of block logs, from December 2004 through December 2024, constituting more than 20 million blocks of users or Ips, in order to examine how blocking, as a core moderation tool, has changed over two decades — time which saw Wikipedia shift to become a more serious, quality-focused enterprise administered by a shrinking community. This study uses an inductive, descriptive approach, aiming to surface normative and procedural changes over time by examining blocks along three dimensions: overall trends, block reasons, and block durations.

Although this study focuses on content moderation, it treats blocking as a primary site where Wikipedia’s governance is enacted. Blocks translate rules-in-form into rules-in-use through a sociotechnical system that includes administrator discretion and technical infrastructure. In addition to contributing to literature about content moderation on Wikipedia, this study extends research about Wikipedia governance in two ways. First, by analyzing the full block log, it shifts rules and norms-based Wikipedia scholarship from policy pages and case studies to the organizational practice of exclusion at scale. Second, by focusing on one form of moderation it provides insight into overall shifting priorities and values of the Wikipedia community amid volunteer scarcity and increased popularity, regardless of whether they are due to changes to rules-in-form or rules-in-use.

Method

Wikipedia is built on MediaWiki software, which automatically maintains logs for a variety of actions.[10] Since 2004, this has included a block log, a record of actions by administrators which prevent users from editing or prevent editing from certain IP addresses. Before the 2004 software update, there were records of blocks, but they were less standardized and never migrated into the new system. For this study, I queried the database using Quarry, a browser

[10] https://www.mediawiki.org/wiki/MediaWiki

interface for SQL queries against the Wikipedia database,[11] and retrieved all logged blocks from the start of the current log on December 23, 2004, through the end of 2024. Data includes the timestamp (when the block was performed), ID (a unique identifier for the action), actor name (the administrator enacting the block), log title (the username or IP address blocked), parameters (the length and scope of the block), and comment text (the reason provided). In addition to blocks, the log contains reblocks (a change in the parameters of an active block) and unblocks, which were omitted due to the scope of this study.

To address RQ1-RQ3, I structured the analysis around three features of the block logs: temporal patterns in overall blocking activity (RQ1), textual rationales provided by administrators (RQ2), and block duration parameters (RQ3).

Logging has undergone several technical and normative changes, with no retroactive standardization. As a result, there is considerable variation in block parameters, including a change to PHP serialized arrays in 2015.[12] While there are tools for administrators with boilerplate rationales (the "comment_text" field of the block log), they are fully customizable. There are furthermore multiple Wikipedia policies which might apply to any particular case. For example, a user inserting unsourced claims in a biography might be blocked for "edit warring," "disruption," or "persistently adding unsourced content to a BLP [biography of a living person]." Blocks for vandalism might cite the vandalism policy in full ("Wikipedia:Vandalism"), one of several understood abbreviations, or any other custom string having to do with the Wikipedia concept of vandalism. Many rationales have typos or mistakes, use shorthand, or reference multiple policies, guidelines, and essays for justification. Some are surprisingly casual or exasperated, like a spammer blocked with the rationale "Just go away."

To understand trends in blocking and reasons for blocking, it was necessary to clean and cluster block rationale data. Each cluster combines multiple keywords referencing related behaviors, typically in reference to particular policies, guidelines, or essays. For example, edit warring is when a user repeatedly makes the same edit or reverts changes made by other users on the same page. [13] The "edit warring" cluster includes rationales with the text "edit warring," "editwar," "three-revert rule," and "3RR," the latter a strict policy limiting the number of reverts any user can make to a page in a 24 hour period; the "promotion" cluster includes "WP:COI" (the conflict of interest guideline)[14] and "paid editing" (a violation of the rule about users paid to edit articles). These are not necessarily synonymous but directly related in terms of the type of behaviors they reference.

Preliminary clustering was performed through a content analysis of referenced policy, guideline, and essay pages,[15] and producing and analyzing a list of the most frequently used block rationales and word frequencies in rationales to manually create a taxonomy. Clusters were

[11] https://quarry.wmcloud.org/

[12] https://www.mediawiki.org/wiki/Manual:Logging_table

[13] https://en.wikipedia.org/wiki/Wikipedia:Edit_warring

[14] https://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest

[15] https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

refined by producing additional word frequencies of blocks not captured by any cluster. In some cases, a single rationale can include reasons covered by multiple clusters, in which case all of the matching clusters are included and reported on (for example, "username contains personal attacks" would be included in both the "username" and "attacks" clusters).

A large "other" category remains due to the custom nature of the field, especially in early years before tools and interfaces with boilerplates were fully developed (although by the time of the earliest blocks in the block log, Wikipedia had been operating for three years, establishing some basic expectations for block reasons). Still, many of the earliest blocks in the log simply use words like "abuse" and "troll," without additional context, although these were common enough to constitute their own cluster. Though also used later, subsequent use was typically combined with other reasons, like "abuse of multiple accounts." Table 1 provides a list of clusters and related keywords.

Cluster name	**Description**	**Keywords**
Abuse	Generic description of a violation, often used in conjunction with other clusters	Abuse
Anonblock	Shared IP addresses blocked due to some misconduct	Anonblock, anon block
Arbitration	Actions by the Arbitration Committee or using rules set forth by the Arbitration Committee (an elected group which functions like a Supreme Court of Wikipedia)	Arbcom, arbitration, WP:AE, WP:ARB
Attacks	Attacking, insulting, or harassing other users	Personal attack, attacking, WP:NPA, harass, insult, civil, battleground, WP:BATTLE, oversight, threat
Compromised	Accounts believed to be compromised, hacked, or for which the owner lost the password	Compromised, hacked, password
Content	Egregious edits to articles or patterns of harmful edits	BLP, biographies, persistent addition, repeated addition, keeps adding, unsourced, uncited, citing sources, verifiability, attack page, nonsense, upload, image, NPOV, WP:POV, POV push, unconstructive
Copyright	Copyright violations	Copyright, copyvio

Disrupt	Vague term for a range of problematic editing behaviors, referencing a guideline called "Disruptive editing."	Disrupt
Editfilter	Any of several problematic behaviors flagged by an automated filtering system	Deliberately triggering, repeatedly triggering
Editwar	Repeatedly making the same edits or repeatedly reverting other users' edits on the same page	Edit war, editwar, edit-war, warring, revert war, revertwar, revert-war, WP:EW, 1RR, 2RR, 3RR, 4RR, 5RR, three revert, three-revert
Legalthreats	Making legal threats	Legal threats, legalthreats, threatening legal, NLT
Nothere	User on Wikipedia for reasons incompatible with the site's goals, a reference the essay "Here to build an encyclopedia."	Nothere, not here to build, not here to collaborate, clearly not here
Other	All reasons not covered by other categories	
Promotion	Abuse of Wikipedia for promotion, undisclosed paid editing, and improper editing with a conflict of interest	Promo, WP:PAID, undisclosed paid, Spam, WP:COI, conflict of interest, advertising, paid edit
Proxy	Blocks of open proxies	Proxy, proxies, webhost, ProcseeBot, ST47ProxyBot, CDNblock, {{tor}}, torblock, exit node, tor node, tor block, vpn
Ranges	IP address ranges blocked due to some misconduct	Schoolblock, school block, rangeblock, range block
Socks	Abusively using multiple accounts	Sock, puppet, checkuser, multiple account, evasion, evade, long-term abuse, long term abuse, longterm abuse, LTA
Troll	Generic term referring to bad faith edits, but can also be used in conjunction with other clusters	Troll
Username	Usernames that are offensive, imitations, promotional, or otherwise violate the username policy	Username, several template names used to address various kinds of username violations, e.g. Uw-uhblock
Vandalism	Deliberately destructive edits	Vand, blanking, hoax, vaublock

Table 1: Block clusters, descriptions, and related keywords

Block duration data was extracted from the parameters field. Unfortunately, this, too, required extensive cleaning as a range of duration types are accepted. Most durations are logged in days, weeks, months, or years, but some specify other relative time periods or end dates. A few administrators seemed to enjoy the occasional block measured in fortnights. Some durations just say e.g. "Friday 6pm," "next Sunday," or ambiguous terms like "3 mon," "now," "next month BST," "fourth day," or "q hours." End dates, where used, do not have a standard format. After several passes to normalize data by number of days, there remained 28,689 blocks with unparsed parameters, some of which do not include a duration and some include it in an obscure format. Technically, there is no difference between "indefinite" and "infinite" blocks; the difference is largely social — there is unlikely to be a reason to lift an infinite block, while an indefinite block is intended to communicate that an unblock is possible in the future if certain conditions are met. These blocks with no expiration complicate operations when calculating trends in block duration, so are represented with a duration of 999 days. This value is greater than any fixed-length block, allowing indefinite blocks to remain part of the statistical summaries, but it should be understood as a placeholder rather than a literal duration. While most of the descriptions below involve medians, the limitations of this convenience value become apparent in mean calculations. Separate means excluding indefinite blocks are not reported, since indefinite cases do not differ in function from fixed-term blocks and excluding them would risk giving the impression of a categorical distinction where none exists.

In addition to my role as an academic researcher, I am also a long-time volunteer contributor to Wikipedia. My familiarity with Wikipedia's policy environment, technical infrastructure, and community norms inform both my access to and interpretation of the data presented here. While this study benefits from contextual knowledge, I have structured the analysis with reproducible methods and clear coding criteria.

Results

Total Blocks

To answer RQ1 (changes in blocking activity over time), I first examined the total number of blocks per year. There was a total of 20,548,435 blocks logged between 2004-2024. Figure 7 shows the total number of blocks per year.

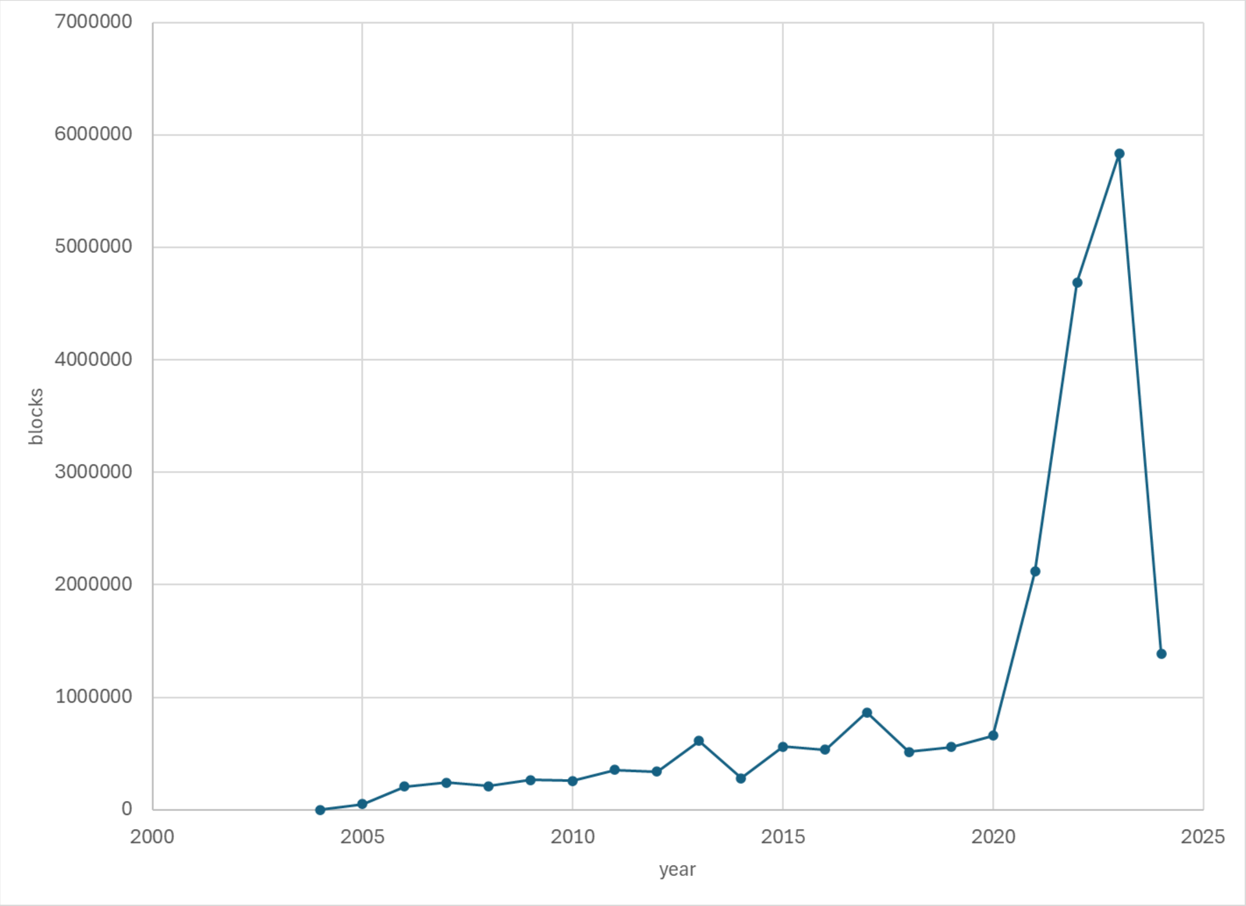

Figure 7: Total blocks per year

From 2004 to 2020 there is a gradual upward trend in block numbers, followed by a significant increase in 2021, peaking in 2023. On inspection, the reason for the sharp increase is due to an effort to preemptively block IP addresses of known open proxies. Using a proxy to edit is generally disallowed on Wikipedia, with rare exceptions.[16] A variety of organizations and individuals compile lists of known proxies, often with unclear methods. In 2009, an automated account (bot) called ProcseeBot[17] made between 91,742 and 667,650 blocks per year from 2009 until it was deactivated by its owner in March 2020. Another bot was created in April 2020, ST47ProxyBot,[18] using additional detection techniques. By the time it was deactivated in March 2024 it made 13,679,025 blocks, accounting for 66.57% of blocks ever made on the English Wikipedia. Together, ST47ProxyBot and ProcseeBot account for 83.55%.

To better understand blocking trends, blocks in the dominant proxy cluster were removed from the rest of the analysis below, yielding a subset of 3,034,268 blocks. Figure 8 reproduces the chart of blocks by year with proxy blocks removed, and both counts are included in Table 2.

[16] https://en.wikipedia.org/wiki/Wikipedia:Open_proxies
[17] https://en.wikipedia.org/wiki/User:ProcseeBot
[18] https://en.wikipedia.org/wiki/User:ST47ProxyBot

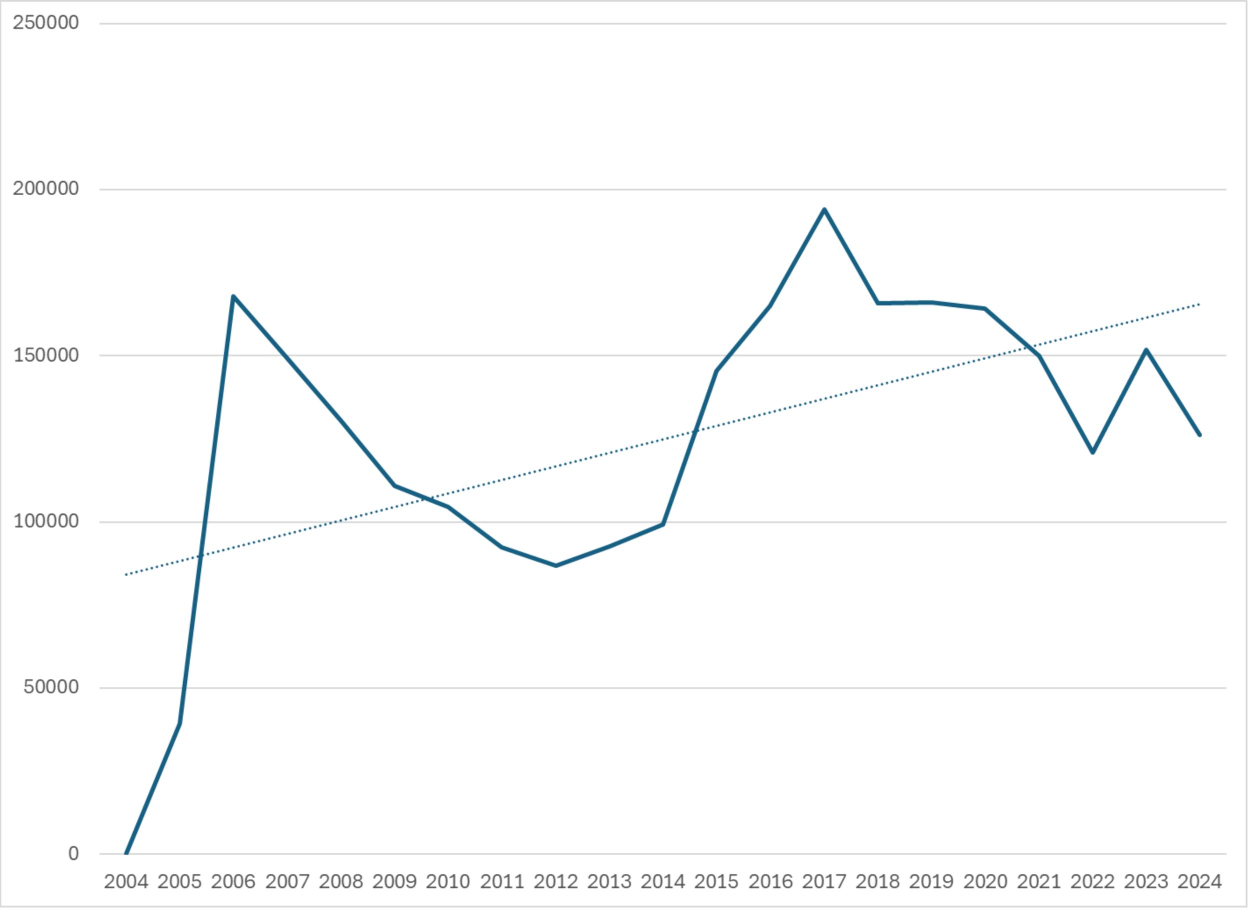

Figure 8: Total blocks per year, other than proxy blocks

Year	Total blocks	Blocks - proxy blocks
2004	375	372
2005	51,577	45,039
2006	206,100	192,248
2007	242,792	221,917
2008	212,228	195,389
2009	264,269	162,140
2010	259,150	151,224
2011	356,516	130,419
2012	339,517	123,653
2013	613,990	123,655
2014	279,920	134,235
2015	562,909	152,643
2016	533,823	165,058
2017	865,866	193,931
2018	515,042	166,137
2019	556,454	164,048
2020	659,388	162,875
2021	2,120,042	150,045
2022	4,688,470	121,321
2023	5,831,555	155,279
2024	1,388,388	122,640

Table 2: Total blocks per year and total blocks other than proxy blocks

Since 2020, users can be blocked from specific pages or namespaces, [19] but this feature is not frequently used. From 2020-2024, there were 6,493 page-specific blocks and 930 namespace blocks. Surprisingly, the frequency has not steadily increased: there were 1,112, 1,548, 1,770, 659, and 1,404 page-level block, and 151, 224, 214, 103, and 238 namespace blocks per year. Further research could assess the impact of partial blocks, but for this analysis I do not distinguish them from sitewide blocks.

[19] https://en.wikipedia.org/wiki/Wikipedia:Partial_blocks

Block reasons

To address RQ2 (changes in administrators' rationales), I clustered the “comment_text” field into policy-based and behavior-based categories and examined their distribution over time. Table 3 lists the block reason clusters and respective counts. Aside from "proxy," which is excluded from subsequent analysis, and "other," reasons are not mutually exclusive. For example, "spammer abusing multiple accounts" would be categorized as both "promotion" and "socks." After proxies, vandalism was by far the most common block rationale, followed by sockpuppetry,[20] username violations, and promotion. "Other" includes all blocks not captured by the defined categories.

Block reason cluster name	**Number of blocks**
Proxy	17,514,167
Vandalism	1,060,214
Username	560,064
Socks	539,096
Promotion	487,866
Abuse	258,377
Ranges	163,067
Disrupt	145,811
Anonblock	132,374
Other	101,470
Content	80,576
Attacks	63,951
Nothere	57,048
Editwar	46,980
Editfilter	24,568
Troll	17,105
Copyright	10,183
Legalthreats	3,956
Arbitration	3,602

[20] https://en.wikipedia.org/wiki/Wikipedia:Sockpuppetry

Compromised	1,692

Table 3: Total blocks by block cluster

Trends emerge when examining reasons over time. Vandalism constituted 63.71% of all blocks in 2004, dropping to between 21-31% since 2014. Username violations have been 15-25% of blocks since 2009. Promotion-related blocks, initially rare, have made up between 15-25% of blocks since 2012. Sock puppetry consistently ranked high, never falling below 7% and peaking at 28.95% in 2024 (see Figure 9). "Nothere" became more common after 2014, reaching 4-5%. Finally, disruption blocks were less than 5% until 2016, growing to 13.53% in 2024. These three trends — disruption, nothere, and other — are highlighted in Figure 10. Table 4 provides a full list of cluster counts by year.

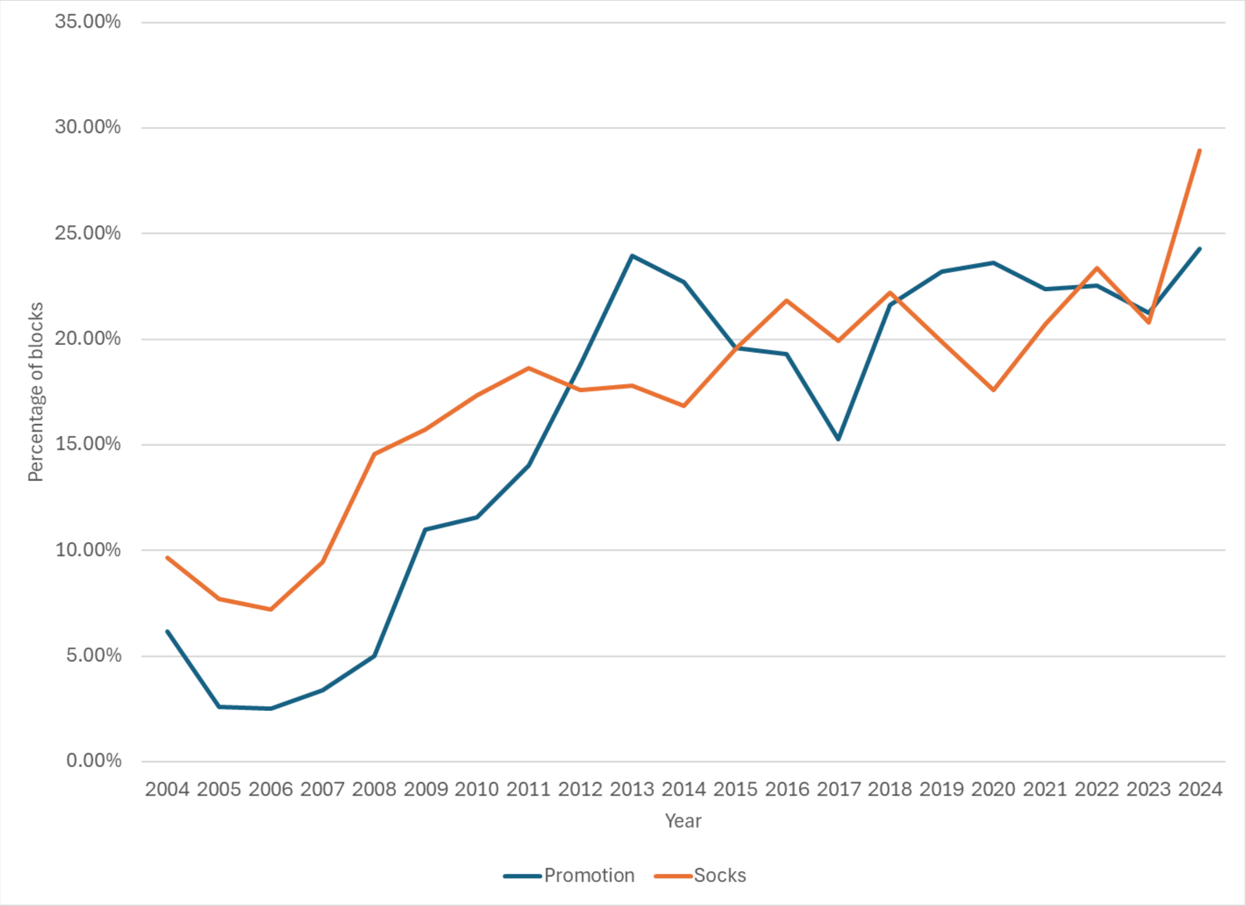

Figure 9: Blocks for promotion and sock puppetry as a proportion of all blocks over time

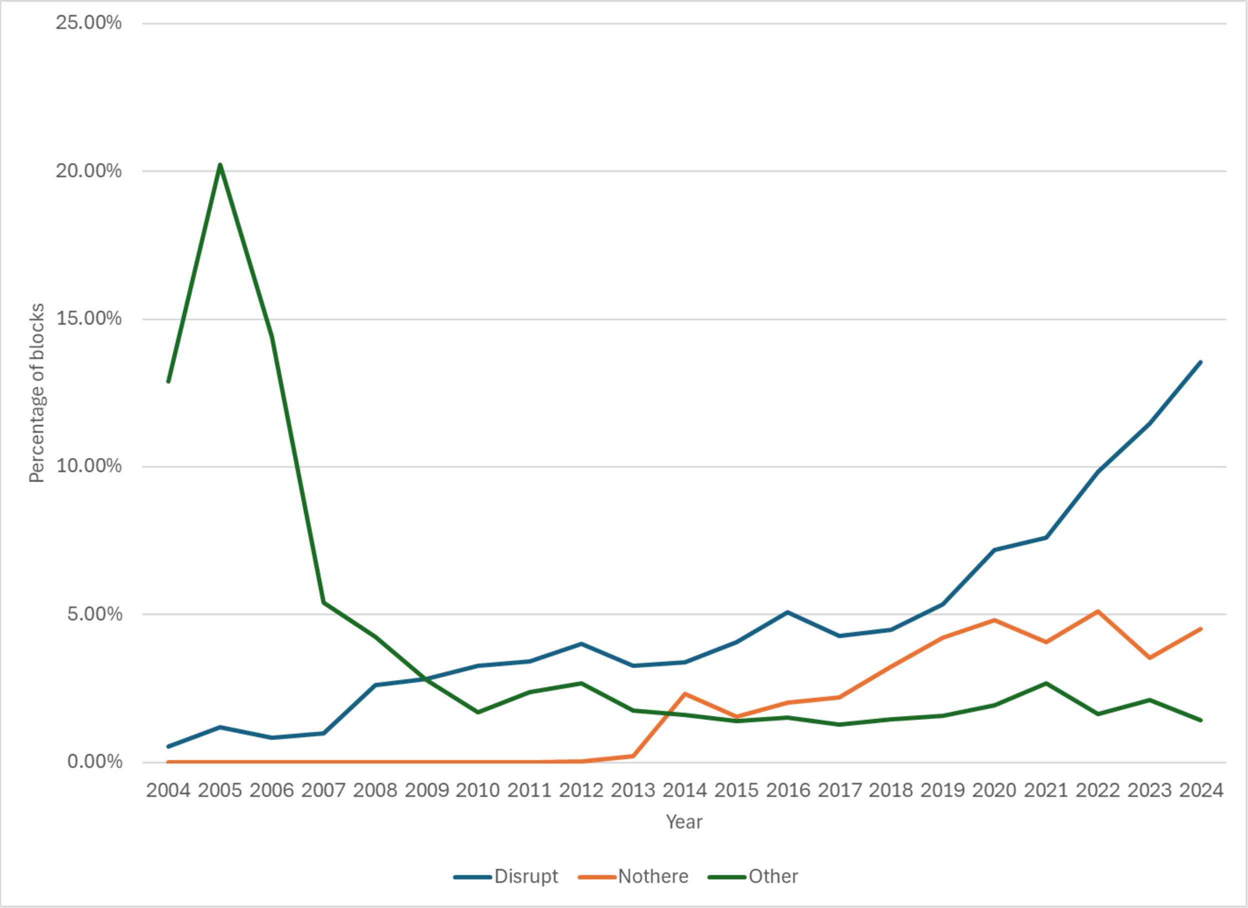

Figure 10: Disruption, nothere, and other block reasons as a proportion of all blocks over time.

cluster	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**
abuse	14	651	2213	9175	23233	19788	17955	16514	14441	10914	10363	11108	13483	16931	14746	13506	11631	13593	12736	12028	13354
anonblock	0	0	100	3890	5646	4878	3961	4501	7082	8462	13244	14505	13922	15768	11040	6254	3884	3917	4097	4137	3086
arbitration	0	357	349	286	260	225	241	189	175	135	134	142	165	117	147	92	103	115	64	128	178
attacks	28	1484	5928	7839	6053	5214	5487	4068	2906	1948	1845	2952	2367	2667	1812	1713	2110	1890	1779	2123	1738
compromised	0	37	97	152	112	94	72	96	65	52	55	55	48	71	122	77	84	79	223	60	41
content	9	976	2530	4946	2788	5245	3968	3334	2871	2677	2361	3153	3902	4739	5060	6426	6672	7840	3873	4302	2904
copyright	6	171	780	948	410	662	630	617	541	427	489	478	582	555	442	463	473	433	367	389	320
disrupt	2	542	1633	2207	5126	4574	4976	4485	4959	4067	4547	6234	8406	8293	7482	8794	11730	11438	*11929*	17791	16596
editfilter	0	0	0	0	0	413	572	678	869	214	194	130	417	7671	620	3096	3519	1637	1273	1282	1983
editwar	40	1613	4088	3919	3637	3059	2427	2458	2345	2235	2315	2332	1730	1873	1641	1475	2088	2014	1876	2012	1803
legalthreats	1	111	285	260	221	280	234	211	233	140	178	185	173	183	163	186	199	209	171	152	181
nothere	0	1	2	19	15	12	19	10	41	257	3132	2394	3371	4257	5386	6942	7839	6101	6222	5490	5538
other	48	9108	27703	12044	8288	4539	2578	3107	3318	2181	2165	2130	2499	2480	2449	2575	3162	4039	2010	3275	1772
promotion	23	1161	4812	7542	9800	17832	17496	18304	23230	29623	30506	29917	31863	29597	35899	38087	38486	33572	27338	32969	29809
ranges	0	6	439	9069	10168	8389	8451	7527	9615	8809	12084	17290	17509	20007	13362	8055	2787	2724	2711	2515	1550
socks	36	3469	13844	20953	28452	25508	26213	24321	21755	22004	22623	29829	36015	38631	36919	32637	28661	31058	28342	32320	35506
troll	7	411	2637	5198	1212	456	454	361	376	338	388	620	713	535	325	372	748	482	575	450	447
username	7	3580	27682	29271	25132	29258	30911	28493	24310	30600	32616	32719	31663	33071	*35130*	31673	33767	29022	23907	24678	22574
vandalism	237	25833	110681	130754	101932	71340	62801	45481	36411	30929	28787	33913	36587	51392	40587	46290	50443	44261	30213	53250	28092

Table 4: Blocks by cluster and year. Top three clusters by year highlighted in cyan; top three years by cluster in italics, highlighted yellow (overlap highlighted green).

Block duration

To address RQ3 (changes in block durations and their implications), I normalized duration data and analyzed trends in median and mean block lengths on their own and across categories. Treating indefinite/infinite blocks as 999 days, the mean block duration is 542.20 days. The median is 999. Both the mean and the median increased from a low mean of 374.03 and low median of 2 in 2005. Since 2010, the median has been indefinite, while the mean has remained between 542-642. Figure 11 charts these trends.

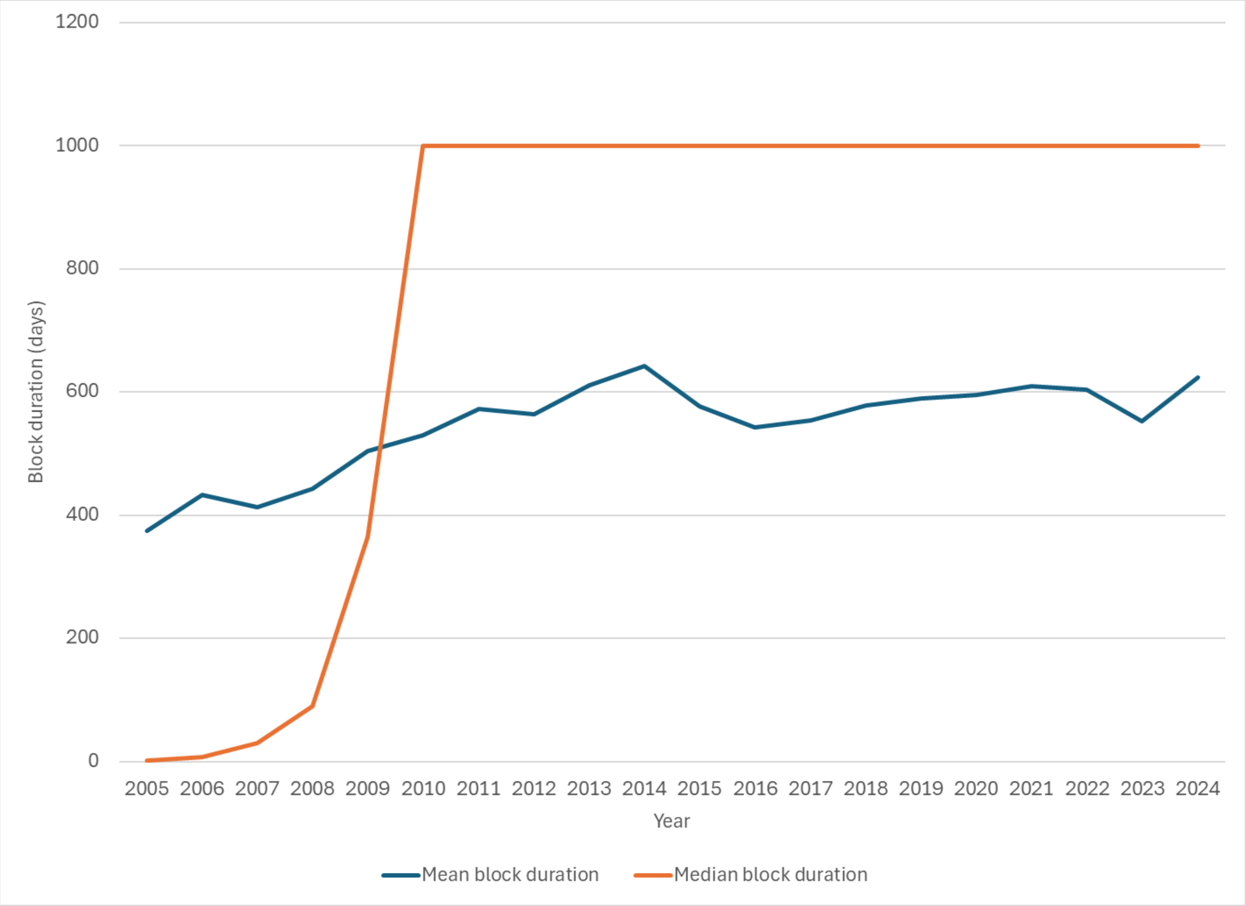

Figure 11: Mean and median block duration (in days) since 2005.

Severity of blockable offenses varies considerably according to the block reason. For example, while sock puppetry is serious enough to almost always receive an indefinite block, edit warring is a relatively minor violation and is more likely to receive a short block. An examination of median block durations shows that while several block rationale clusters were consistent, copyright, disruption, edit warring, and, to a lesser extent, vandalism all show upward trends. In other words, blocks tend to be longer for these offenses in recent years compared to years past, while none of the reasons had pronounced downward trends. See Table 5 for a full breakdown.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	all years
abuse	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
anonblock		1.65	30	30	30	30	30	30	30	30	30	90	90	90	90	30	14	14	30	90	30
arbitration	90	14	365	7	14	7	14	14	30	11	14	7	7	7	7	7	14	60	14	7	14
attacks	2	7	30	7	30	7	7	7	7	14	7	7	7	3	7	14	30	14	14	14	7
compromised	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
content	1	7	90	999	14	30	7	7	7	14	7	7	7	90	999	7	14	14	7	14	14
copyright	2	2	7	7	7	14	14	14	30	999	90	999	999	999	999	999	999	999	999	999	365
disrupt	2	2	7	2	3	2.29	3	7	7	7	2.50	2	2	3	3	7	7	7	90	90	7
editfilter					14	1.29	1.29	1.29	1.29	1.29	1.29	1.29	999	30	1.29	7	1.29	1.29	1.29	1.50	2
editwar	1	1	1	1	1	1.29	1.29	1.29	1.50	2	2	1.50	1.29	1.50	1.50	2.50	3	7	7	7	1.29
legalthreats	30	30	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
nothere	0.13	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
other	999	999	999	999	999	180	999	999	999	999	999	7	3	7	3	30	999	30	999	840	999
promotion	1	4.13	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
ranges	0.01	90	90	117	180	180	180	180	180	180	90	90	90	90	90	180	365	365	180	180	90
socks	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
troll	999	999	999	999	999	999	999	999	999	999	30	999	90	30	999	999	999	999	999	999	999
username	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
vandalism	1	1.29	2	2.29	3	7	7	14	30	999	999	14	90	14	14	30	30	7	1.29	14	7

Table 5: Median block duration by reason and year.

Labor

Because blocking patterns relate to changes in administrative capacity, I additionally examined how blocking labor was distributed among administrators over time. Since December 2004, 2,092 administrators have made blocks on the English Wikipedia (2,088 excluding proxy blocks). The number of admins making at least one block per year peaked in 2007 at 1,103 before declining to a low of 415 in 2024. Mean blocks per admin has seen an overall upward trend, peaking in 2017. Figure 12 shows the total number of admins making blocks in each year since 2004 and the mean number of blocks per admin in the same time period.

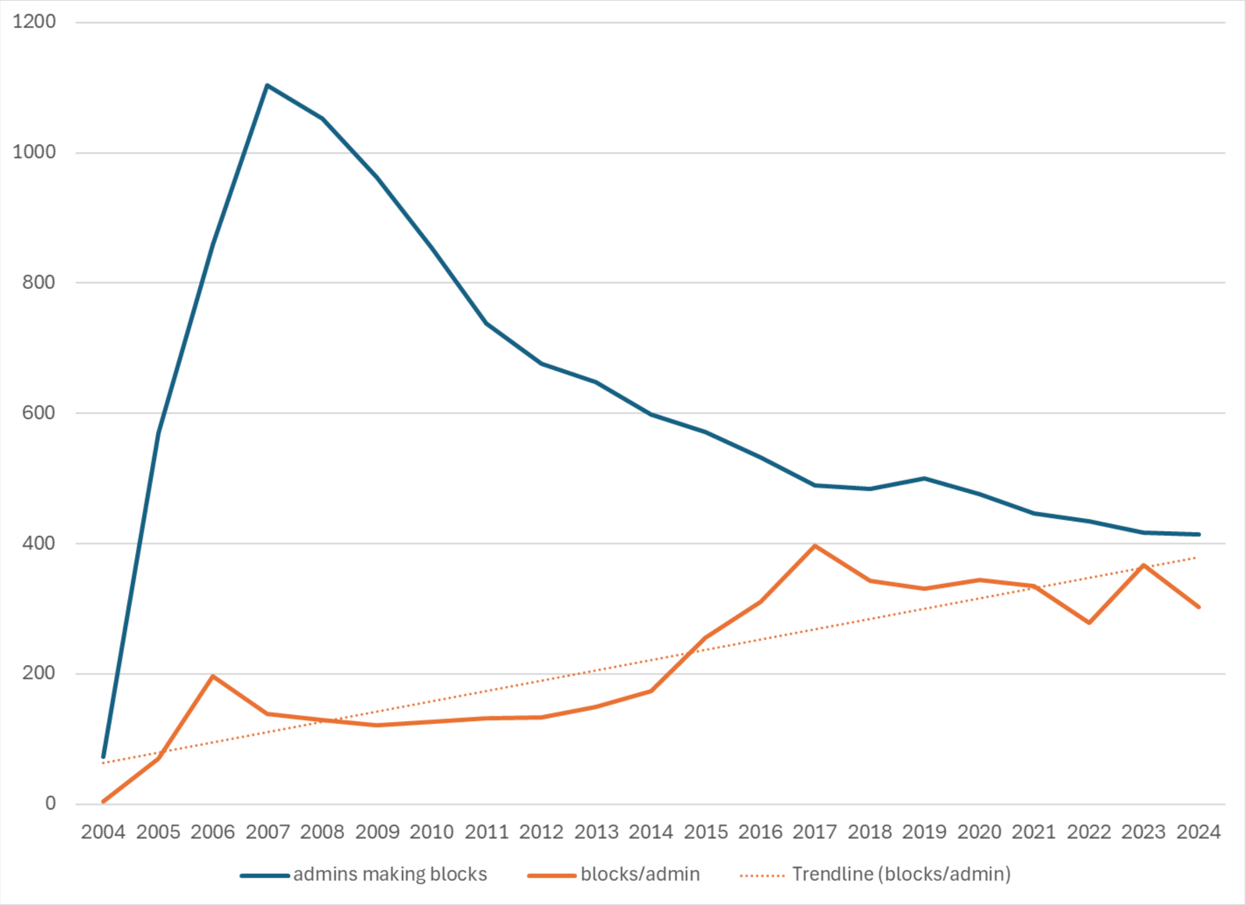

Figure 12: Total admins making blocks each year and mean blocks per admin.

Figure 13 illustrates the percentage of blocks performed by the most active 10% of administrators per year. In general, after beginning at 49.19% in 2004, blocking work was more concentrated in the most active 10%, peaking at 84.98% in 2016 and hovering between 75-85% from 2009-2023. In 2024, however, the most active admins made only 37.22% of blocks. The precise reasons for this sudden drop are unclear, but there is at least one highly active administrator, User:Materialscientist, whose activity level dropped significantly from 2023 to

2024. This editor alone has made more than three three times as many blocks as any other administrator, and their activity dropped from 38,965 blocks in 2023 to 5,707 blocks in 2024.

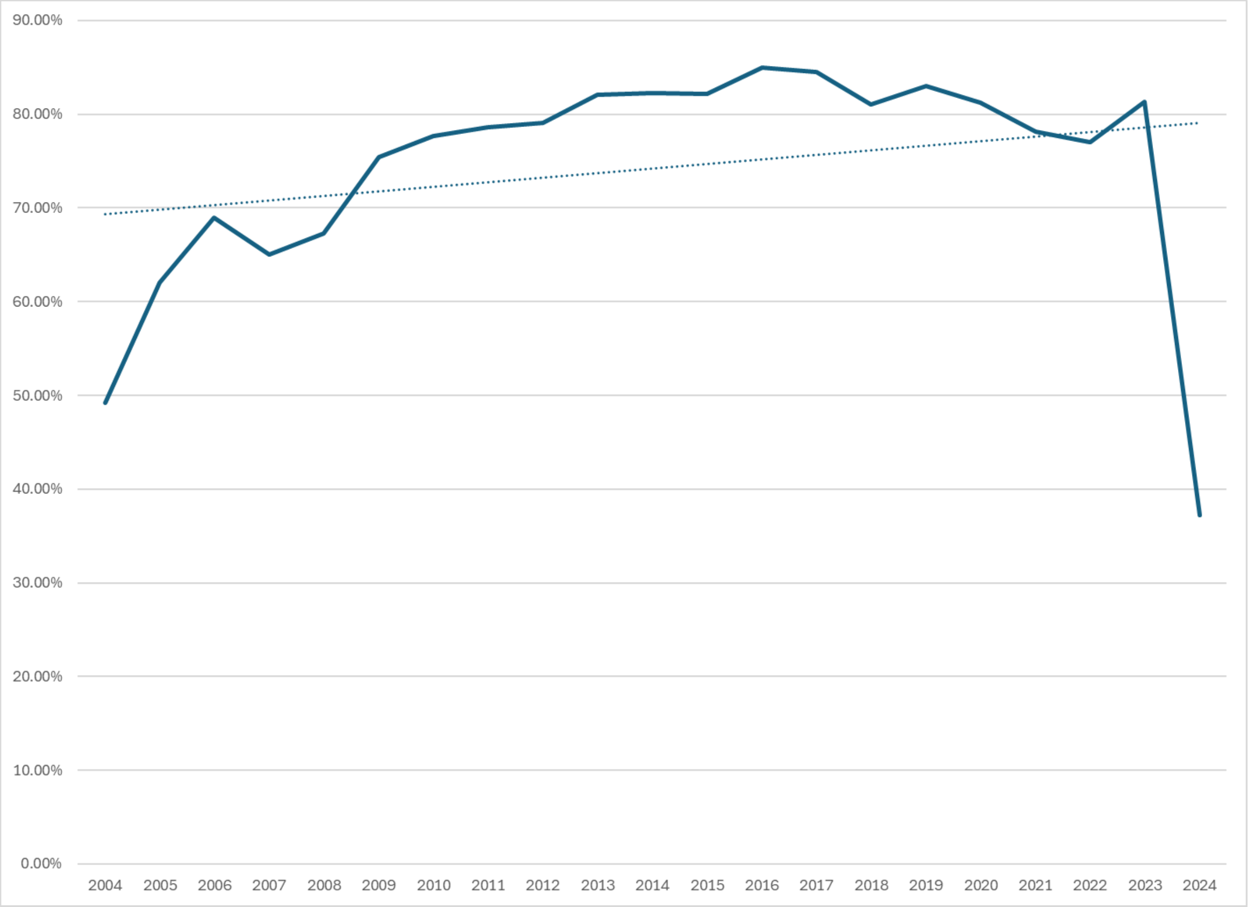

Figure 13: Blocks by the most active 10% of administrators per year.

Discussion

Examination of two decades of blocking data shows more preemptive blocks, longer durations, broader justifications, and a concentration of blocking work among a shrinking pool of administrators. The importance of Wikipedia as a source of knowledge in the world is not lost on contributors, who have steadily tightened standards for quality on the site. The block log reveals how administrators scale volunteer labor to maintain standards amid ever-increasing size and popularity. None of this "boundary-work" is objectively desirable or undesirable but reflects trade-offs between openness and either quality or the need to manage limited labor.

The number of admins making blocks has steadily decreased since a high in 2007, and the mean number of blocks per admin has gradually increased. The work is not evenly distributed, with a smaller number of admins making most of the blocks. 10% of admins typically make more than 80% of the blocks, although these numbers are heavily skewed by a small number of highly active admins. These especially active administrators are likely more difficult to recruit should the current power admins retire.

Unlike for-profit platforms, Wikipedia cannot simply use money to recruit more labor. Recruiting administrators requires attracting new users, and indeed the possibility of future productivity is the primary disincentive to blocking, but that possibility must be weighed against the cost of constructive volunteers' time to fix future errors. It is a calculation that may implicitly preference more, rather than less, restriction where a pattern of positive work has not yet been demonstrated (Halfaker et al., 2012).

These findings extend existing research on the evolution of Wikipedia's rules-based governance. While rules-in-use are not so unstable as to undermine collaboration (Keegan & Fiesler, 2017), many of the trends visible in the block log are not rooted in changes to codified policy but normative shifts corresponding to broader – and difficult to measure -- attitudes, or even the result of the adoption of technical tools. While Wikipedia's rules and norms like "assume good faith" were initially credited for making its openness possible in the first place (Reagle, 2010), their increased breadth, complexity, and enforcement have been criticized for creating an unwelcoming atmosphere for new contributors (Halfaker et al., 2012). To the extent stricter enforcement is a result of a shrinking pool of labor, it may also contribute to the same problem.

Ethnographic and other qualitative researchers have shown that Wikipedia's editor community is socially heterogeneous, with contributors discursively constructing the Wikipedia "community" in ways that involve their own personal motivations (Pentzold, 2010). Fians' analysis of edit wars in articles on British royals illustrates the way good faith contributors can come into conflict over contested narratives (2024). "Good faith" and "bad faith," then, are functional administrative terms rather than sociological descriptions of contributors' personal motivations.

The most conspicuous change is the rise of preemptive blocks. While proxies have long been disallowed due to abuse, scaling up the practice of finding and acting upon lists of known proxies began in response to prolific vandals using proxies to vandalize many pages, taking a lot of volunteer time to clean up. A single proxy-blocking bot accounts for most of the blocks in the history of the English Wikipedia. Preemptive proxy blocking is an excellent example of the trade-off between openness, productivity, and costs in Grimmelmann's content moderation taxonomy. IP addresses can be reassigned, requiring frequent testing to ensure an IP blocked as a proxy is not reassigned to a normal internet user. Furthermore, the practice of blocking proxies in general — preemptive or otherwise — is controversial. For example, Tran, et al. found that Tor users (on IPs that had not yet been blocked) contributed content roughly on par with other unregistered or new users (2019). Whether due to state censorship or harassment, there are many good reasons to permit contributors a greater degree of anonymity.

The reasons provided by administrators for blocks have changed over time, likely due to changes in the challenges faced by Wikipedia, policy interpretations, or administrative norms. Promotional activity and sockpuppetry grew as a proportion of all blocks, which makes sense considering Wikipedia's importance in the world — there are more people interested in influencing the site now compared to twenty years ago. In the earliest parts of the log, the "other" rationale cluster was larger and reasons often had either simple but imprecise language like "troll" or highly specific narrative reasons. As time went on, the project's policies and guidelines became more developed and administrators had additional tools they could use to simplify the blocking practice, including boilerplates and templates. The “other” rationale shrank while more generalized and standardized rationales became more common.

Two examples are “nothere” and “disruption.” The former is a reference to an essay, "Here to build an encyclopedia," which interprets policies like "neutral point of view" and "assume good faith" to distinguish users by intentions. It is an incredibly broad rationale, summarized in the page’s “nutshell” banner as "Wikipedians are here to build an encyclopedia, i.e., a neutral, reliable public reference work on notable topics. Users whose behavior suggests they are here for some other purpose risk being blocked or banned".[21] While it may not go far enough in acknowledging the gray area between good faith and bad faith, it provides a pragmatic basis for necessarily subjective administrative decisions. In 2013, it was added to a list of options administrators could choose from in a drop-down menu in the MediaWiki interface.[22] Nothere blocks jumped from 41 blocks in 2012 to 3,132 in 2014 and 7,838 in 2020. Adding a justification to the interface means people will choose that justification, and it was added to the interface because it is useful to have a broad justification covering a wide swath of bad faith actors. Blocks for disruption have increased steadily to the point of being among the most common in recent years. The guideline page about disruption uses similarly vague language: "a pattern of editing that disrupts progress toward improving an article or building the encyclopedia".[23] As with nothere blocks, its usage increased after it was added to the drop-down menu in 2007, but continued to increase thereafter. Standardization can be beneficial to new user experience in the way it ensures that block rationales are tied to concrete policies or guidelines, typically linked from the rationale itself, although there is limited value in such directions if the rationales are overly broad. A trend towards more vague explanations contravenes existing best practices for content moderation, whereby clarity and perceptions of fairness are related to sanction acceptance and rehabilitation (Jhaver, et al., 2018; Chang & Danescu-Niculescu-Mizil, 2019).

There was an upward trend in median block durations for copyright, disruption, edit warring, and vandalism, but no similarly obvious downward trends. Since 2009, there are more indefinite blocks than fixed-length blocks overall, and mean block durations have increased gradually over time. Increasing block durations may be due to an increased proportion of bad faith new users, but is more likely a symptom of the need to scale (blocking for longer to avoid the need for additional moderation decisions in the future) or of tightening standards (providing fewer chances to make mistakes in order to protect content). In either case, it falls short of the ideal of graduated sanctions for commons-based moderation (Ostrom, 1990), increasing the likelihood that the blocked user will see the action as unfair and abandon the project (Chang & Danescu-Niculescu-Mizil, 2019). It may be worth the community's effort to investigate alternative

[21] https://en.wikipedia.org/wiki/Wikipedia:Here_to_build_an_encyclopedia
[22] https://en.wikipedia.org/wiki/MediaWiki:Ipbreason-dropdown
[23] https://en.wikipedia.org/wiki/Wikipedia:Disruptive_editing

interventions, such as the tool proposed by Halfaker et al. which flags not just bad faith new users but good faith new users (2014), or strategies to apply "procedural justice" to social platforms (e.g. Badiei et al., 2020).

There are several limitations on this work, the most obvious being the reliance on logs and policy pages. Future work could focus on blocked users, interview administrators, and take a random sample of blocks to examine closely. There is relatively little work comparing blocking with less severe moderation tools could be used, such as those that impose a “design friction” (Chandrasekharan, et al., 2022) like page protection. This study also did not examine changes to block settings ("reblocks") or unblocks, which could shed light on, for example, which users or which offenses are more likely to have blocks overturned. Topic modeling tools might also aid understanding of how block rationales, which I group manually, relate to one another. Most significantly, it is limited to the English Wikipedia. Each Wikipedia is developed separately, with often similar but parallel governance systems and policies and norms around blocking that vary according to local attitudes, strategies, and capacities.

Conclusion

Over the past two decades, Wikipedia has evolved from a small, open experiment to a serious global knowledge resource. In tracing two decades of the English Wikipedia's blocking practices, this study has examined how a volunteer-run, non-profit, sociotechnical system has changed one of its core mechanisms of content moderation amid growing external demands and shrinking internal capacity. Analysis of more than 20 million block log entries between 2004-2024 shows how blocking activity (RQ1), rationales (RQ2), and durations (RQ3) have shifted: overall blocking activity showed a gradual upward trend apart from a sharp recent increase in preemptive proxy blocking; block rationales became more standardized, with a tendency towards vaguer categories like “disruption;” and median block durations increased across several categories, including disruption, edit warring, copyright, and vandalism.

These changes reflect evolving norms about what behaviors threaten the project, how administrators justify exclusion, and movement towards longer-term preventative measures. Wikipedia’s volunteers have had to develop strategies to “do more with less,” tightening standards and renegotiating the boundaries of acceptable participation as the project’s size and visibility expanded. Moderation decisions reflect a balance between the principles of openness Wikipedia was founded on and long-term resilience, as administrators navigate limited labor and increasing expectations for quality.

Beyond Wikipedia, these findings connect to broader debates in content moderation. While Wikipedia’s governance is unusually transparent and volunteer-driven (Jemielniak, 2013; Tkacz, 2015), the patterns observed in this study, such as preemptive intervention, longer sanctions, and vaguer justifications, echo themes seen in research on moderation of commercial platforms (Gillespie, 2018; Roberts, 2019; Gorwa, et al., 2020). The underlying logics are quite different, but broader content moderation literature suggests that pressures of scale, reputational risk, and labor availability can shape moderation systems even without factoring in the pursuit of profit, suggesting there may be an opportunity for more cross-platform research that considers not just platforms' rules-in-form but rules-in-use (Keegan & Fiesler, 2017), technological affordances, and labor constraints. Future research should therefore not only examine how blocked users on Wikipedia and comparable projects experience and interpret these sanctions, how administrators and moderators apply evolving norms in practice, and how blocking interacts with other moderation tools, but also compare these dynamics across non-profit, volunteer-run, and commercial platforms. Wikipedia has adapted remarkably well to the pressures of its own success, maintaining and in many cases increasing quality beyond what many thought possible, but its adaptations, and their analogs on other platforms, are not without an as-yet poorly understood cost.

Bibliography

Asikin-Garmager E, Liou Y, Lo C, Myrick C, Gerdemann B and Chen D (January 2025). Wikipedia Administrator Recruitment, Retention, and Attrition. *WMF Research.* Available at: https://upload.wikimedia.org/wikipedia/commons/5/5f/%28Final_Report%29_Administrator_recruitment%2C_retention%2C_%26_attrition_%28SDS1.2.2%29.pdf (accessed April 16, 2025).

Badiei F, Meares T and Tyler T (2020) Community Vitality as a Theory of Governance for Online Interaction. *Yale Journal of Law & Technology, 23*.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D (2020) Language Models Are Few-Shot Learners. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 159: 1877-1901.

Butler B, Joyce E and Pike J (2008) Don't Look Now, But We've Created a Bureaucracy: the Nature and Roles of Policies and Rules in Wikipedia. *CHI '08: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 1101-1110. DOI: https://doi.org/10.1145/1357054.1357227

Chandrasekharan E, Jhaver S, Bruckman A and Gilbert E (2022) Quarantined! Examining the Effects of a Community-Wide Moderation Inervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI), 29*(4): 1-26. DOI: https://doi.org/10.1145/3490499 https://doi.org/10.1145/3490499

Chang J and Danescu-Niculescu-Mizil C (2019) Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. *WWW '19: The World Wide Web Conference*: 184-195. DOI: https://doi.org/10.1145/3308558.3313638

Choi B, Alexander K, Kraut RE and Levine JM (2010) Socialization Tactics in Wikipedia and Their Effects. *CSCW 2010*. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=cf55a99a17b66f1cf47cadcc142ba8913b1ce3ef (accessed April 16, 2025).

Fians, G. (2024). The Death of Elizabeth II on Wikipedia: Fleshing Out Freedom Through Technoliberal Participation Online. *Journal of the Royal Anthropological Institute, 30*(4): 912-931. DOI: 10.1111/1467-9655.14115

Ford H (2022) *Writing the Revolution: Wikipedia and the Survival of Facts in the Digital Age*. Cambridge, MA: MIT Press.

Geiger RS (2009) The Social Roles of Bots and Assisted Editing Tools. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration..* New York: ACM Digital Library. Available at: http://www.stuartgeiger.com/papers/geiger-wikisym-bots.pdf (accessed April 16, 2025).

Geiger RS (2011) The Lives of Bots. In G Lovink and N Tkacz (Eds.) *Wikipedia: A Critical Point of View*. Amsterdam: Institute of Network Cultures.

Gieryn TF (1983) Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review, 48*(6): 781-795.

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

Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*. DOI: 10.1177/2053951719897945.

Grimmelmann J (2015) The Virtues of Moderation. *Yale Journal of Law & Technology, 42*. Available at: https://scholarship.law.cornell.edu/facpub/1486/ (accessed April 16, 2025).

Halfaker A, Geiger RS, Morgan JT and Riedl J (2012) The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity is Causing Its Decline. *American Behavioral Scientist, 57*(5): 664-688. DOI: 10.1177/0002764212469365.

Halfaker A, Geiger RS and Terveen LG (2014) Snuggle: Designing for Efficient Socialization and Ideological Critique. *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 311-320. DOI: 10.1145/2556288.2557313.

Jankowski S (2023) The Wikipedia Imaginaire: a New Media History Beyond Wikipedia.org (2001-2022). *Internet Histories, 7*(4): 333-353. DOI: https://doi.org/10.1080/24701475.2023.2246261

Jemielniak D (2014) *Common Knowledge? An Ethnography of Wikipedia*. Redwood City, CA: Stanford University Press.

Jhaver S, Ghosal S, Bruckman A and Gilbert E (2018) Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction, 25*(2): 1-33. DOI: https://doi.org/10.1145/318559

Jiang J and Vetter M (2019, August 14) The Good, the Bot, and the Ugly: Problematic Information and Critical Media Literacy in the Postdigital Era. *Postdigital Science and Education, 2*: 78-94. DOI: 10.1007/s42438-019-00069-4.

John N (2024) A Classification of Features for Interpersonal Disconnectivity in Digital Media: Block, Unfriend, Unfollow, Mute, Withhold, and Eject. *Media and Communication, 12,* Article 8716. DOI: https://doi.org/10.17645/mac.8716

Keegan B and Fiesler C (2017) The Evolution and Consequences of Peer Producing Wikipedia's Rules. *Eleventh International AAAI Conference on Web and Social Media, 11*(1): 112-121. DOI: https://doi.org/10.1609/icwsm.v11i1.14899

Kittur A, Suh B, Pendleton BA and Chi EH (2007) He Says, She Says: Conflict and Coordination in Wikipedia. *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 453-462. DOI: https://doi.org/10.1145/1240624.1240698

Kriplean T, Beschastnikh I, McDonald DW and Golder SA (2007) Community, Consensus, Coercion, Control: cs*w or How Policy Mediates Mass Participation. *GROUP '07: Proceedings of the 2007 ACM International Conference on Supporting Group Work*: 167-176. DOI: https://doi.org/10.1145/1316624.131664

McDowell ZJ and Vetter MA (2020) It Takes a Village to Combat a Fake News Army: Wikipedia's Community and Policies for Information Literacy. *Social Media + Society 6*(3). DOI: 10.1177/2056305120937309.

McDowell Z and Vetter M (2022) Fast "Truths" and Slow Knowledge; Oracular Answers and Wikipedia's Epistemology. *Fast Capitalism*: 104–112. DOI: 10.32855/fcapital.202201.009.

McGrady R (2009) Gaming against the greater good. *First Monday, 14*(2). DOI: 10.5210/fm.v14i2.2215.

McGrady R (2013) Ethos [edit]: Procedural Rhetoric and the Wikipedia Project. In M. Folk & S. Apostel (Eds.), *Online Credibility and Digital Ethos: Evaluating Computer Mediated Communication.* IGI-Global. DOI: 10.4018/978-1-4666-2663-8.

Miller C, Smith M, Marsh O, Balint K, Inskip C and Visser F (2022) Information Warfare and Wikipedia. *Institute for Strategic Dialogue.* Available at: https://www.isdglobal.org/wp-content/uploads/2022/10/Information-Warfare-and-Wikipedia.pdf (accessed September 25, 2025).

Morgan JT and Halfaker A (2018) Evaluating the impact of the Wikipedia Teahouse on newcomer socialization and retention. In: *Proceedings of the 14th International Symposium on Open Collaboratio*n, Paris France, 22 August 2018, pp. 1–7. ACM. DOI: 10.1145/3233391.3233544.

Ostrom E (1990) *Governing the Commons: the Evolution of Institutions for Collective Action.* Cambridge, UK: Cambridge University Press.

Pentzold, C. (2010). Imagining the Wikipedia Community: What Do Wikipedia Authors Mean When They Write About Their 'Community'? *New Media & Society, 13*(5). DOI: 10.1177/1461444810378364

Reagle J (2010) *Good Faith Collaboration: The Culture of Wikipedia.* Cambridge, Massachusetts: MIT Press.

Rijshouwer E, Uitermark J and De Koster W (2023) Wikipedia: a Self-Organizing Bureaucracy. *Information, Communication, & Society, 26*(7): 1285-1302. DOI: 10.1080/1369118X.2021.1994633.

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media.* New London: Yale University Press.

Schiff S (2006, July 23) Know It All. *The New Yorker.* Available at: https://www.newyorker.com/magazine/2006/07/31/know-it-all (accessed September 25, 2025).

Schluger C, Chang JP, Danescu-Nicolescu-Mizil C and Levy K (2022) Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2): 1-27. DOI: https://doi.org/10.1145/3555095

Shah-Quinn N (2025, January 24) Investigating the 2021–present registration decline. *Wikimedia Foundation*. Available at: https://people.wikimedia.org/~nshahquinn-wmf/2021-registration-decline/ (accessed March 6, 2025).

Tkacz N (2015) *Wikipedia and the Politics of Openness*. Chicago: University of Chicago Press.

Tran C, Champion K, Forte A, Hill BM and Greenstadt R (Are Anonymity-Seekers Just Like Everybody Else? An Analysis of Contributions to Wikipedia from Tor. *2020 IEEE Symposium on Security and Privacy*: 186-202. DOI: https://doi.org/10.1109/SP40000.2020.00053

Wikimedia Foundation (n.d.) *Wikimedia Statistics*. Available at: https://stats.wikimedia.org/#/en.wikipedia.org (accessed March 5, 2025).

Wulczyn E, Thain N and Dixon L (2017) Ex Machina: Personal Attacks Seen at Scale. *WWW'17: Proceedings of the 26th International Conference on World Wide Web*: 1391-1399. DOI: https://doi.org/10.1145/3038912.305259