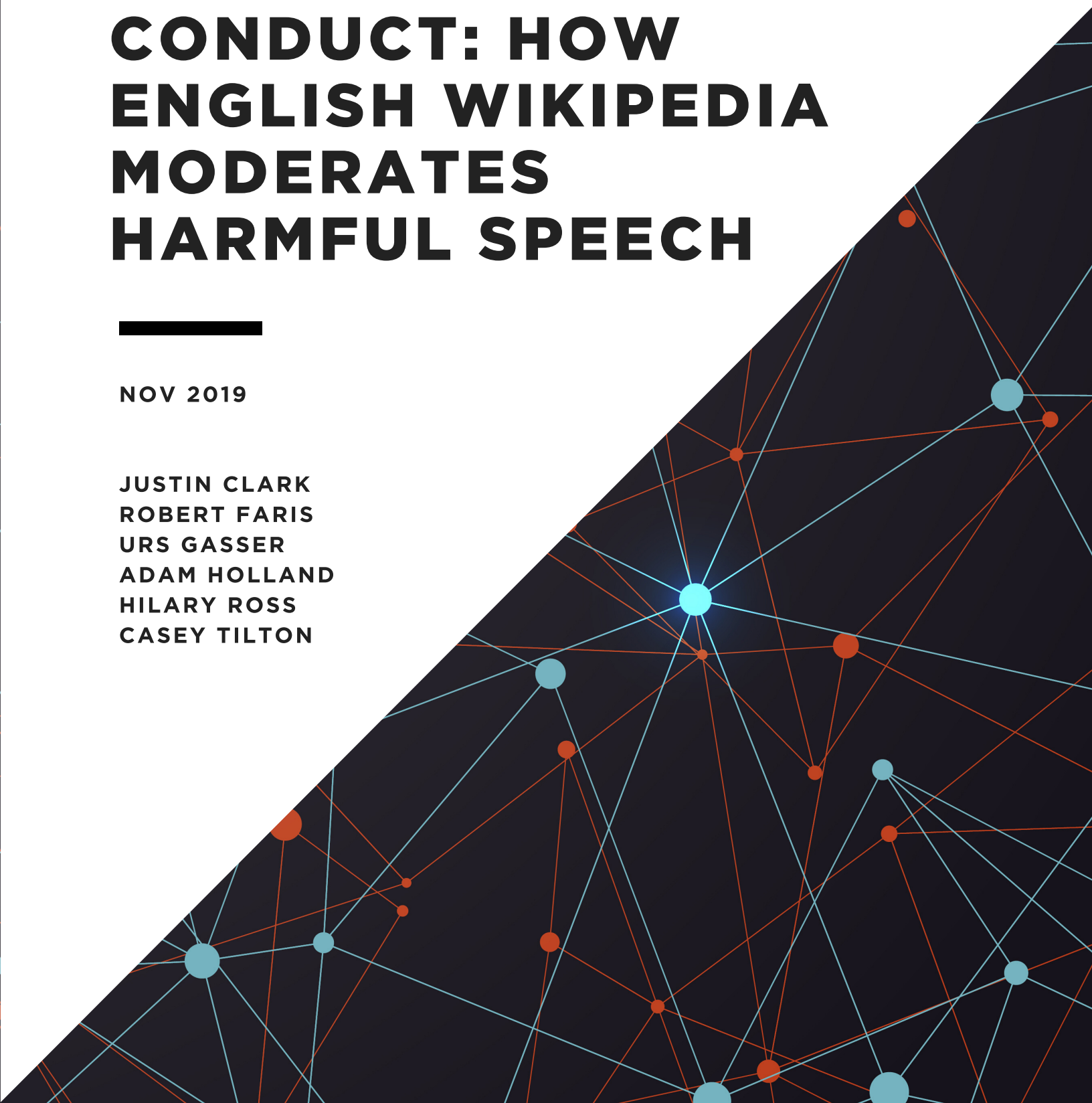
Content and Conduct: How English Wikipedia Moderates Harmful Speech

BERKMAN
KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

CONTENT AND CONDUCT: HOW ENGLISH WIKIPEDIA MODERATES HARMFUL SPEECH

NOV 2019

JUSTIN CLARK
ROBERT FARIS
URS GASSER
ADAM HOLLAND
HILARY ROSS
CASEY TILTON



Acknowledgments

The authors would like to thank the following people for their critical contributions and support: The 16 interview participants for volunteering their time and knowledge about Wikipedia content moderation; Isaac Johnson, Jonathan Morgan, and Leila Zia for their guidance with the quantitative research methodology; Patrick Earley for his guidance and assistance with recruiting volunteer Wikipedians to be interviewed; Amar Ashar, Chinmayi Arun, and SJ Klein for their input and feedback throughout the study; Jan Gerlach and other members of the Legal and Research teams at the Wikimedia Foundation for their guidance in scoping the study and for providing feedback on drafts of the report; and the Wikimedia Foundation for its financial support.

Executive Summary

In this study, we aim to assess the degree to which English-language Wikipedia is successful in addressing harmful speech with a particular focus on the removal of deleterious content. We have conducted qualitative interviews with Wikipedians and carried out a text analysis using machine learning classifiers trained to identify several variations of problematic speech. Overall, we conclude that Wikipedia is largely successful at identifying and quickly removing a vast majority of harmful content despite the large scale of the project. The evidence suggests that efforts to remove malicious content are faster and more effective on Wikipedia articles compared to removal efforts on article talk and user talk pages.

Over time, Wikipedia has developed a multi-layered system for discovering and addressing harmful speech. The system relies most heavily on active patrols of editors looking for damaging content and is complemented by automated tools. This combination of tools and human review of new text enables Wikipedia to quickly take down the most obvious and egregious cases of harmful speech. The Wikipedia approach is decentralized and defers much of the judgements about what is permissible or not to its many editors and administrators. Unlike some social media platforms, which strive to clearly articulate what speech is acceptable or not, Wikipedia has no concise summary of what is acceptable and not. Empowering individuals to make judgements about content, which extends to content and conduct that some would deem harmful, naturally leads to variation across editors in the way that Wikipedia's guidelines and policies are interpreted and implemented. The general consensus among those editors interviewed for this project was that Wikipedia's many editors make different judgements about addressing harmful speech. The interviewees also generally agreed that efforts to enforce greater uniformity in editorial choices would not be fruitful.

To understand the prevalence and modalities of harmful speech on Wikipedia, it is important to recognize that Wikipedia is comprised of several distinct, parallel regimes. The governance structures and standards that shape the decisions of Wikipedia articles are significantly different from the processes that govern behavior on article talk pages, user pages, and user talk pages.

Compared to talk pages, Wikipedia articles receive a majority of the negative attention from vandals and trolls. They are also the most carefully monitored. The very nature of the encyclopedia-building enterprise and the intensity of vigilance employed to fight vandals who target articles means that Wikipedia is very effective at removing harmful speech from articles. Talk pages, the publicly viewable but behind-the-scenes discussion pages where editors debate changes, are often the location of heated and bitter debates over what belongs in Wikipedia.

Removal of harmful content from articles, particularly when done quickly, is most likely a deterrent for bad behavior. It is also an effective remedy for the problem. If seen by very few readers, the fleeting presence of damaging content results in proportionately small harm. On the other hand, the personal attacks on other Wikipedians—frequently on talk pages—is not as easily mitigated. Taking down the offending content is helpful but does not entirely erase the negative impact on the individual and community. Governing discourse among Wikipedians continues to be a major challenge for Wikipedia and one not fixed by content removal alone.

This study, which builds upon prior research and tools, focuses on the removal of harmful content. A related and arguably more consequential question is the effectiveness of efforts to inhibit and prevent harmful speech from occurring on the platform in the first place by dissuading this behavior ex ante rather than mopping up after the fact. Future research could adopt similar tools and approaches to this report to track the incidence of harmful speech over time and to test the effectiveness of different interventions to foster pro-social behavior.

The sophisticated and multi-layer mechanisms for addressing harmful speech developed and employed by Wikipedia that we describe in this report arguably represent the most successful large-scale effort to moderate harmful content, despite the ongoing challenges and innumerable contentious debates that shape the system and outcomes we observe. The approaches and strategies employed by Wikipedia provide valuable lessons to other community-reliant online platforms. Broader applicability is limited as the decentralized decision-making and emergent norms of Wikipedia constrain the transferability of these approaches and lessons to commercial platforms.

Overview and Background

Harmful speech online has emerged as one of the principal impediments to creating a healthy, inclusive Internet that can benefit all. Harmful speech has at times been viewed as an unavoidable by-product of democratizing communication via the Internet. Over the past several years, there is a growing appreciation that minority and vulnerable communities tend to bear the brunt of online harassment and attacks and that this constitutes a major obstacle to their fully participating in economic, social, and cultural life online.

The problems of governing conduct online on large platforms such as Facebook, Twitter, and YouTube are widely recognized. The multiple challenges include the sheer size of the effort and attention required to moderate content at the scale of large platforms, which is coupled with the limitations of automated tools to moderate content; the expertise, judgement, and local knowledge required to understand and address harmful speech in different cultural contexts and across different jurisdictions; the difficulty in crafting robust moderation guidelines and policies that can be sensibly replicated by tens of thousands of moderators; the psychological burden borne by content moderators that spend their work days sifting through highly disturbing content; the highly dynamic political pressures being applied; and the unprecedented concentration of power over permissible speech in a small number of private companies.[1]

There are a number of prior studies that have focused specifically on Wikipedia's content moderation practices and policies. R. Stuart Geiger and David Ribes examine the process of "vandal fighting" on Wikipedia, emphasizing the complex interactions between editors, software, and databases.[2] The paper

[1] Key scholarship in this area includes Gillespie, T. (2018). Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media. New Haven: Yale University Press. http://custodiansoftheinternet.org/. Archived at https://perma.cc/LV6Z-XNN4; Keller, Daphne (2019). Who Do You Sue? State and Platform Hybrid Power over Online Speech, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1902. https://www.lawfareblog.com/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech. Archived at https://perma.cc/S84F-MQYV; Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. Harvard Law Review, 131(6), 598-670. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/. Archived at https://perma.cc/R43U-D67Z; Marwick, A. E., & Miller, R. (2014). Online harassment, defamation, and hateful speech: A primer of the legal landscape. *Fordham Center on Law and Information Policy Report*, (2). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2447904; Citron, Danielle Keats, & Mitchell, Stephen A. (2014). Hate Crimes in Cyberspace. Harvard University Press. https://www.daniellecitron.com/hate-crimes-in-cyberspace/. Archived at https://perma.cc/RV7T-69F4; Roberts, Sarah T. (2014). Behind the Screen: Content Moderation in the Shadows of Social Media. https://yalebooks.yale.edu/book/9780300235883/behind-screen. Archived at https://perma.cc/8HP5-93AP;
Caplan, Robyn. Context or Content Moderation. https://datasociety.net/output/content-or-context-moderation/. Archived at https://perma.cc/25SG-K4TE; Kaye, D. (2019). Speech police : The global struggle to govern the Internet. New York: Columbia Global Reports. https://globalreports.columbia.edu/books/speech-police/. Archived at https://perma.cc/3HTB-SPB4; Arun, Chinmayi. "Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms" https://medium.com/berkman-klein-center/rebalancing-regulation-of-speech-hyper-local-content-on-global-web-based-platforms-1-386d65d86e32. Archived at https://perma.cc/5EWS-RF38.

[2] Geiger, R. S., & Ribes, D. (2010, February). The work of sustaining order in wikipedia: the banning of a vandal. In Proceedings of the 2010 ACM conference on Computer supported cooperative work (pp. 117-126). ACM.

highlights the importance of editor collaboration through technological tools, including bots, scripts, and assisted editing programs, and it concludes that these tools fundamentally transformed "the moral order" and practice of editing on Wikipedia.

Whereas Geiger and Ribes focus on technological tools, Paul de Laat investigates Wikipedia's content moderation bureaucracy.[3] He examines how Wikipedia editors conceive of their participation, and how this leads them to perceive rules on Wikipedia. To do this, he reviews the history of Wikipedia's governance up to 2012, details the process for reviewing edits, and lastly, analyzes debates over proposals to introduce a system of review for new edits on Wikipedia. He concludes that, "Wikipedians either rejected the proposed scheme (because it is antithetical to their conception of Wikipedia as a community) or endorsed it (because it is consonant with their conception of Wikipedia as an organization with clearly defined boundaries)." De Laat's work shows the importance of understanding where policies come from and how editors perceive their contributions to Wikipedia, as these factor into whether and how new practices and policies are adopted.

In "The Virtues of Moderation," James Grimmelmann offers a taxonomy and overview of moderation in online communities with a case study on the "sophisticated and intricate" structures and practices of content moderation on Wikipedia.[4] He stresses the fact that content moderation relies as much on social norms as technology and is hence "always emergent, contingent and contestable." He also highlights the tensions between the decentralized organization and strong reliance on social norms on Wikipedia that shape content moderation on the platform.

Dariusz Jamielniak's *Common Knowledge?: An Ethnography of Wikipedia* is a comprehensive ethnographic study of the platform and its users. It attempts to solve the puzzle of how and why Wikipedia's unique organizational design works. The book describes the role that conflicts and dissent play in article development and explores the importance of user status and peer control despite the community's anti-hierarchical stance. It also highlights how the community embraces pseudonymity and rejects credential checking by placing trust in procedures rather than trust in the expertise of individuals.[5]

Nathan Matias explores the various frames used to describe the work of moderators on Reddit, which has many parallels to Wikipedia. Considering three ways in which moderators may conceive of their work —digital labor, civic participation, and oligarchy among moderators—he finds that they adopt aspects of these frames and introduces the term civil labor to encompass the complex, multifaceted roles that volunteer moderators play.[6]

In order to improve policies and tools, the Wikimedia Foundation has done substantial work to understand the state of harassment on Wikipedia. In 2015, Wikimedia's Support & Safety team conducted a multi-language survey on harassment. The survey defined harassment as: name calling, trolling or flaming, content vandalism, stalking, outing or doxxing, discrimination, impersonation,

[3] de Laat, P. B. (2012). Coercion or empowerment? Moderation of content in Wikipedia as essentially contested bureaucratic rules. Ethics and information technology, 14(2), 123-135.

[4] James Grimmelmann, *The Virtues of Moderation*, 17 Yale J.L. & Tech (2015).

[5] Jemielniak, Dariusz. Common Knowledge?: An Ethnography of Wikipedia. Stanford University Press, 2014.

[6] Matias, J. N. (2019). The civic labor of volunteer moderators online. Social Media + Society, 5(2).

revenge porn, threats of violence, or other.[7] Results indicated that 38% of respondents could confidently recognize they'd been harassed on Wikipedia, and 51% had witnessed others being harassed. Content vandalism and trolling were the most frequent forms of harassment. When asked for suggestions about how to improve the Wikimedia movement, respondents named improving Wiki governance, improving reporting mechanisms, improving culture, and technical solutions as some of the top preferred potential solutions.

In part, this led to a research collaboration between Wikimedia and Jigsaw to better understand the nature and impact of harassment on Wikipedia and to develop technical solutions. In investigating personal attacks, one form of harmful content, researchers discovered that only 18% of personal attacks were moderated (attacking users getting warned or blocked) and that 67% of attacks on English Wikipedia were made by registered users.[8]

The core objective of this study is to understand, document, and evaluate efforts on English-language Wikipedia to moderate harmful speech. Employing two complementary approaches—interviews with Wikipedians and content analysis—we find that Wikipedia is quite successful in quickly removing the vast majority of harmful speech despite the large volume of content revisions. As we show in this report, Wikipedia is most effective in moderating harmful content on the articles of the platform and somewhat less effective in policing talk pages. This conclusion is bolstered both by the general consensus of those that we interviewed and the quantitative content analysis. For the quantitative analysis piece, we started by estimating how much content is removed from English Wikipedia through the various removal methods. We then estimated how much content could qualify as harmful speech using a fairly narrow definition. Using the identified pieces of harmful content, we estimated the average lifetime of harmful speech in various English Wikipedia namespaces and characterize the community of editors who remove that speech.

We describe and document in this study the systems and mechanisms that Wikipedia employs to identify and remove harmful speech. This analysis is informed in large part by the interviews with Wikipedians that we conducted for the study. We interviewed 16 Wikipedia editors about the processes and guidelines for content revision, content deletion, and quality control of English Wikipedia. The interviews focused on gaining an understanding of the community's policies and decision making about how to handle harmful content both on articles and talk pages. The focus of the conversations were not on the actions of individual editors and administrators but rather to understand how the system operates overall. Appendix 1 includes a list of interview questions and extended summaries of the responses to each question.

To help ground and guide the analysis, we explore the complexities of defining harmful speech, offer a taxonomy that highlights the many overlapping forms of harmful speech, and summarize the ways in

[7] Harassment Survey 2015. Wikimedia Foundation. https://upload.wikimedia.org/wikipedia/commons/5/52/Harassment_Survey_2015_-_Results_Report.pdf. Archived at https://perma.cc/B88W-3YV6.

[8] Wulczyn, E., Thain, N., & Dixon, L. (2017, April). Ex machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web (pp. 1391-1399). International World Wide Web Conferences Steering Committee.

which different forms of harmful speech map against the guidelines and policies that shape conduct and content on Wikipedia.

As we describe later in further detail, Wikipedia has developed a hybrid system of volunteer editors and automated tools for moderating content that operates across the several namespaces of the platforms, including the articles, user pages, and article talk pages. In effect, Wikipedia simultaneously operates multiple regulatory regimes that employ different sets of tools and have different objectives. One regime guides the actions of its volunteer editors in the creation and maintenance of the encyclopedia. Another mediates the interpersonal conduct of the editors while they do their work. Both operate based on a detailed set of guidelines and policies that are meant to reflect the social norms that have emerged over the many years of the project but that are implemented in a decentralized way, which offers a lot of flexibility and autonomous judgement in their application.

The issues we take on in this study align with broader questions about the governance of online platforms. To help put the approaches and efforts of Wikipedia to police conduct and content into a broader context, we devote a section of this report to a comparison of the different regulatory approaches that have emerged to govern online speech, spanning traditional legal approaches, private ordering, and the coregulatory and hybrid models that occupy the space in between. This helps to position the strategies adopted by Wikipedia within this regulatory landscape.

The unique and well developed mechanisms for addressing harmful speech developed and employed by Wikipedia that we describe in this report arguably represent the most successful large-scale effort to moderate harmful content, despite the ongoing challenges and innumerable contentious debates that shape the system and outcomes we observe. The decentralized decision making and emergent norms of Wikipedia—a hallmark of the project as well as a source of great acrimony and deserved pride—also limit the transferability of these approaches and lessons to commercial platforms.

The next section explores the concept of harmful speech and presents a taxonomy of harmful speech developed for this study.

What is Harmful Speech?

Harmful speech online comes in many forms, from hate speech directed at groups of people based on race or ethnicity, to personal attacks, harassment, defamation, threats, bullying, stalking, or the posting of personal information or intimate photos without consent, etc.[9]

Tracking and quantifying harmful speech is a difficult endeavor. There is no canonical definition for harmful speech and even were it clearly defined, determining what does and does not constitute harmful speech is highly subjective and context specific. Assessing harmful speech on Wikipedia is complicated in that Wikipedia-specific ontologies don't line up easily with the way that academics, other non-profits, companies, and governments frame harmful speech. As we describe in this report, Wikipedia has developed sophisticated systems for identifying and removing a broad range of harmful

[9] https://cyber.harvard.edu/publications/2016/UnderstandingHarmfulSpeech. Archived at https://perma.cc/2LQB-CVTB.

speech on the platform. There is not, however, a simple summary that describes these processes and no simple road map for outsiders to understand how this works.

In this section we present a simple taxonomy that we later employ to bridge different frameworks and ontologies and to map together legal standards of impermissible speech to the concepts and categories used by scholars in the field and the language and processes used by Wikipedia.

To guide and frame our analysis for this study, we focus on five categories of harmful attacks against others that map reasonably well against guidelines and policies on Wikipedia.[10] [11]

Harassment – This category includes typically repeated behaviors that induce personal distress or fear. Examples might include cyberbullying, stalking, or in the Wikipedia context, wikihounding—repeatedly engaging with an editor in order to "confront or inhibit their work."[12]

Harassment is a topic of major concern on Wikipedia and the subject of longstanding efforts at prevention and mitigation.[13]

Threats – This includes threats of physical harm or sexual assault as well as incitement to violence.

Threats are similarly a topic of intense concern for Wikipedia as one of many forms of impermissible personal attacks.[14]

Defamation – This category broadly captures personal attacks that damage one's reputation.

Generally, defamation is a statement that causes injury to the reputation of a third party. It can be spoken (slander) or written (libel). Under United States jurisprudence, a plaintiff seeking to prove defamation must successfully show four separate elements. "1) a false statement purporting to be fact; 2) publication or communication of that statement to a third person; 3) fault amounting to at least negligence; and 4) damages, or some harm caused to the person or entity who is the subject of the statement."[15] Politicians or public figures such as celebrities have an even higher burden, and must also show "actual malice," meaning that the allegedly harmed speaker must prove that the speaker or writer either knew the statement was false or acted with reckless disregard for whether it was true.[16]

[10] For a more detailed and comprehensive typology of content that is illegal in some countries see https://www.internetjurisdiction.net/uploads/pdfs/Papers/Content-Jurisdiction-Program-Operational-Approaches.pdf. Archived at https://perma.cc/LLQ3-X8HL.

[11] Intellectual property infringement, which would constitute another category, falls outside the scope of this project.

[12] https://en.wikipedia.org/wiki/Wikipedia:Harassment. Archived at https://perma.cc/MLP3-PRSL.

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

[14] https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks. Archived at https://perma.cc/8MFA-HFUD.

[15] https://www.law.cornell.edu/wex/defamation. Archived at https://perma.cc/Y85V-63NB.

[16] http://www.dmlp.org/legal-guide/proving-fault-actual-malice-and-negligence. Archived at https://perma.cc/HW94-U7P8.

Biographical articles about living persons are closely scrutinized on Wikipedia and subject to tight guidelines on what is permissible. This is a response in part to concerns over defamation.[17]

Identity-based attacks – Many attacks are based in part on attacking a person's group characteristics, including race, gender, ethnicity, nationality, gender, gender identity, sexual orientation, disabilities, illnesses, or age, among others. This may include attacks that span a broad range of speech that is discriminatory, demeaning, or dehumanizing. This category overlaps substantially with common definitions of hate speech.

On Wikipedia, attacks based on group affiliation is at the top of the list of comments that are "never acceptable": [18]

> Abusive, defamatory, or derogatory phrases based on race, sex, sexual orientation, gender identity, age, religious or political beliefs, disabilities, ethnicity, nationality, etc. directed against another editor or a group of editors. Disagreement over what constitutes a religion, race, sexual orientation, or ethnicity is not a legitimate excuse.

Posting personal information – Revealing sensitive personal information without consent might include posting location, place of work, financial information, health history, personal relationships, non-consensual intimate images, etc. The term "doxxing" is often used in reference to online instantiations of this category. Revealing non-public information about others is a primary concern on Wikipedia and efforts are made to quickly remove such information.

These categories are not mutually exclusive. Many attacks span multiple categories. For example, posting the address of a person and threatening to harm them because of their race covers several categories simultaneously. There are also many other forms of harmful speech that do not fall neatly into this simple framework.

These categories map usefully against legal frameworks across different countries.[19] Threats of violence and incitement to violence are illegal in many jurisdictions. Defamation is a civil matter in many countries and illegal in others. The group-based attacks category broadly overlaps with definitions of hate speech, which is illegal in many countries and notably not illegal in the United States. Legal standards and remedies to the posting of personal information online is an unsettled and emerging area of law. Harassment spans a wide range of conduct, some of which reaches the level of potentially illegal conduct in some jurisdictions. Much of it falls into the awful but lawful realm.

Content that is targeted for removal on Wikipedia does not always map cleanly against legal standards across different countries. There is indeed content that is removed from the platform that is clearly illegal in most countries, such as threats to commit violence and severe forms of harassment. Wikipedia content and conduct policies often go beyond the requirements of the law and target text that is

[17] https://en.wikipedia.org/wiki/Wikipedia:Biographies_of_living_persons. Archived at https://perma.cc/YKJ9-ZDUA.
[18] https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks. Archived at https://perma.cc/8MFA-HFUD.
[19] See the Mandola project for a mapping and summary of what constitutes illegal hate speech in different European countries. http://mandola-project.eu/. Archived at https://perma.cc/9YQR-DSQ6.

potentially legal in many jurisdictions. In this study, we focus attention primarily on attacks that are likely to result in personal harm, and within this area on identity-based attacks, harassment, and threats. Vandalism, a long-standing challenge on Wikipedia, often includes forms of personal attacks that fall within the focus of this study. We do not address wider issues of vandalism not related to attacks that may result in personal harm. Although some would consider disinformation to be a form of harmful speech, we do not address it in this study. Similarly, we do not include in this study policies and actions to address attempts to defraud people online or intellectual property infringement, which for some might fall within the realm of harmful speech.

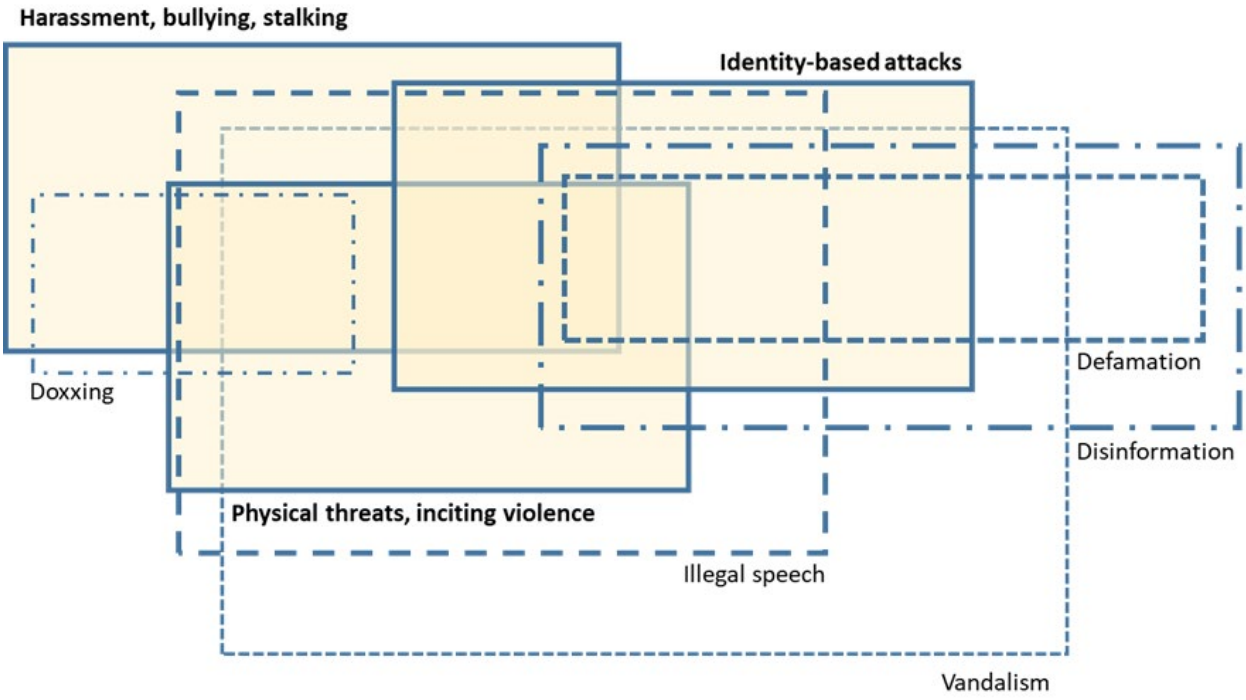


Figure 1. Forms of harmful speech and focal areas for this study

The content classifiers we describe in a later section to estimate the degree to which harmful speech is found on Wikipedia coincide roughly with the identity-based attacks, harassment, and threat categories.

Reducing the scope of inquiry to harmful personal attacks does not mitigate all of the inherent fuzziness in articulating and identifying harmful speech. Conceptualizations of harmful speech might draw on the intent of the speaker, the content itself, or the impact of the speech—did it result in actual harm? None of these conceptual approaches offer simple guidance or structure for identifying and measuring harmful speech. Another complicating factor is that there is no single threshold above which offensive

speech becomes harmful. Identifying and addressing harmful speech, or studying it, means delving into an area that is highly subjective.[20]

Approaches and Models for Governing Content Online

Governing speech, whether on Internet platforms, in schools and businesses, or as a legal matter, follows the same set of basic processes. First, there has to be a set of rules, standards, guidelines, or principles that define the boundary between acceptable and unacceptable speech. It helps if the boundaries are clearly defined, though this is always a challenge given the highly context dependent and subjective nature of speech. Where there are common cultural and normative foundations for interpreting the meaning of speech, those too are helpful. Second, if these constraints are to be enforced, there must be mechanisms to discover unacceptable speech capable of covering the scale of speech. Third, responsibility must be assigned to those who will decide whether speech falls into the acceptable or unacceptable category. Fourth, a remedy for unacceptable speech must be specified and implemented. This might mean legal sanction, or, on privately governed platforms, the takedown of the speech, a warning, or user account suspension. Another useful feature is a process by which determinations that speech is unacceptable can be reviewed, appealed, and reversed if found to be wrongly determined.

The central regulatory questions for the governance of speech on online platforms is how to allocate responsibility for these various roles. Who sets the rules and standards? Who identifies potentially impermissible speech? Who decides whether it is impermissible? Who takes what action? Can that action be appealed? Across different countries and different online platforms, Wikipedia one of them, the answers to these questions vary substantially. There are a modest number of general models that have emerged which we summarize here. This set of general models includes those that are defined by law, those that are privately ordered, and a set of hybrid models. These models represent a wide array of different mechanisms regarding notification and discovery of the content to be removed, actions taken, recourse for platforms and posters, and general transparency. Here, we summarize the most common general models currently in place and attempt to categorize them at a high level.

Models Articulated and Designed in Law

Unsurprisingly, many of the best known models of content removal are the product of explicit law or regulations. Three of the most relevant are: “notice-and-takedown,” the removal of online content that is directly initiated by government actors, and a newer model that we have termed “pre-emptive responsibility” that places more direct responsibility for content (and therefore possible liability) with the platforms themselves.

[20] For a more detailed description of the challenges and different perspectives on defining harmful speech, see https://cyber.harvard.edu/publications/2016/UnderstandingHarmfulSpeech. Archived at https://perma.cc/2LQB-CVTB. Also see Andy Sellars for an excellent overview of perspectives on defining hate speech: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244. Archived at https://perma.cc/39HD-H65E.

Notice-and-Takedown

By far the most common type of content removal model, notice-and-takedown (N&TD), sometimes called "notice and action," follows a pattern whereby a platform has no duties of any kind regarding a piece of content until it receives notice of some kind that the content is violating some law or regulation. Upon receiving notice, the platform must then take some action, whether that be removal or blocking of the material, or something else. Variants include "notice and notice," for example in Canada,[21] or "notice and stay down,"[22] which has not been put into practice yet but is on the wishlist for many large-scale rights-holders[23] and governments.[24]

Perhaps the most well-known example of a notice-and-takedown model is the United States' Digital Millennium Copyright Act, the DMCA.[25] Passed into US Federal law in 1998 as a way to address the burgeoning copyright concerns arising from the rapidly growing popularity and power of the Internet, the DMCA provides a safe harbor from legal liability for copyright infringement if the hosting entity removes the content in question upon being notified of its presence. Many other removal regimes and protocols, both legal and private, are modeled after, or borrow from, the DMCA.

India also has a N&TD + safe harbor schema for online content,[26] with the interesting distinction that platforms are only required to remove content after receiving the "notice" in the form of a court order or other notification directly from a government agency. That is, an individual's complaint directly to the platform need not be acted upon.

Germany's NetzDG law, the "Network Enforcement Act," is also a species of N&TD.[27] In this Act, the German government has placed responsibility on platforms for the removal, after a complaint from a user, of any content that is unlawful within the meaning of various sections[28] of German law and its criminal code. Platforms are responsible for creating easy to use and transparent mechanisms for users to report putatively unlawful content, and must remove any reported content within either a 24-hour or seven-day time frame, depending on the nature of the content. NetzDG is also an interesting model because it has size and traffic thresholds[29] platforms must meet before being subject to the law, and

[21] http://www.ic.gc.ca/eic/site/oca-bc.nsf/eng/ca02920.html. Archived at https://perma.cc/GA6B-R9HR.
[22] https://en.wikipedia.org/wiki/Notice_and_take_down#Notice_and_stay_down. Archived at https://perma.cc/R8P2-LZZF.
[23] https://www.theguardian.com/technology/2016/mar/24/bpi-british-music-labels-piracy-policy-google. Archived at https://perma.cc/Z9M4-498T.
[24] https://www.ccianet.org/2018/03/new-eu-recommendation-on-illegal-content-online-undermines-online-rights-and-harms-europes-tech-economy/. Archived at https://perma.cc/XR4B-CLLP.
[25] https://www.copyright.gov/title17/. Archived at https://perma.cc/763D-B7LK.
[26] For more information on India's N&TD protocols, see Appendix 2
[27] https://germanlawarchive.iuscomp.org/?p=1245. Archived at https://perma.cc/MRY6-WAR8; For more information on the NetzDG law, see Appendix 2.
[28] Id.
[29] Id.

because it has mandated transparency provisions regarding complaints received, provisions that Facebook has already been fined for violating.[30]

Another newer yet well-known N&TD mechanism is found within the EU's General Data Protection Regulation (GDPR). Known in the past and more colloquially as the "right to be forgotten" and also sometimes as the "right of erasure,"[31] this allows EU citizens to request the erasure of their personal data from data controllers and processors such as databases and search engines. No action need be taken by a search engine or other party until a request is made, and once made, the recipient must evaluate the request according to six different criteria. If any one of these six apply, the recipient of the erasure request "shall have the obligation to erase personal data without undue delay," although there are five possible exceptions to this obligation.[32]

At least with respect to search engines, the required geographic scope of such a removal was a point of dispute, with Google and others being willing only to remove search results with respect to the country of origin of the request—in this case the EU—not globally. The French data protection authority, CNIL, demanded that search results be removed globally, regardless of the whereabouts of the viewer or relevant local law. In September of 2019, the EUCJ ruled in favor of Google in Case C-507/17 *Google v CNIL*, holding "there is no obligation under EU law for Google to apply the European right to be forgotten globally and clarifying further that while EU residents have the legal right to be forgotten, the right only applies within the borders of the bloc's 28 Member States."[33]

An October 3, 2019 ruling, *Eva Glawischnig-Piesczek v. Facebook Ireland Limited,*[34] appears to in part contradict the CNIL finding. In this case, the EUCJ ruled against Facebook, holding that after receiving a court order to remove allegedly defamatory material regarding an Austrian politician, Facebook was not only required to do so but had to do so globally. See the following section for more details.

Government-initiated Removal of Illegal Speech

In this type of model, a particular government or legal regime not only defines a category or categories of speech as illegal, it implements a related schema within which anything meeting the criteria for that category of speech must be taken down when an aspect of government, typically a specialized agency but sometimes a court, demands or otherwise seeks to compel that the content be removed. For government requests that do not come through official or formal legal channels, and where compliance is not mandatory, see the section below about informal cooperation.

---

[30] https://www.dw.com/en/germany-fines-facebook-for-underreporting-hate-speech-complaints/a-49447820. Archived at https://perma.cc/Z8RQ-SARF.

[31] https://gdpr-info.eu/art-17-gdpr/. Archived at https://perma.cc/6XAY-CFNN.

[32] Id.

[33] https://europeanlawblog.eu/2019/10/29/google-v-cnil-case-c-507-17-the-territorial-scope-of-the-right-to-be-forgotten-under-eu-law. Archived at https://perma.cc/EP4Q-XDZU.

[34] http://curia.europa.eu/juris/document/document.jsf?text=&docid=218621&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=1965965. Archived at https://perma.cc/VZ8E-DFKR.

The Russian Federal Service for Supervision of Communications, Information Technology, and Mass Media (Roskomnadzor)[35] is perhaps the best known of such agencies, issuing many demands to online platforms such as Twitter or WordPress to remove content having to do with drug use and suicide techniques, among other topics, as well as content violating Russian federal law 276-FZ "On Amendments to the Federal Law 'On Information, Information Technologies and Data Security,'" commonly referred to as the "VPN law." Roskomnadzor will often block access to certain IP address until content is removed.[36]

The Turkish government, typically through subdivisions such as local criminal courts, also sends a substantial volume of orders for the removal for content that is alleged to violate Turkish Law.[37] Notably, the government provides little to no transparency regarding these requests, requiring researchers to rely on collecting them from recipients or from the Lumen[38] project.

The Indian government, relying on the provisions of a few separate laws, has the power to block, filter, or remove content under Section 69A of the IT act, or to disable access to content by disabling access to the Internet itself.[39] See Appendix 2 for more details about a recent example of completely disabling access in the Kashmir region.

More recently, the Singaporean Parliament passed The Protection from Online Falsehoods and Manipulation Act ("POFMA"),[40] which will be administered within the Info-communications Media Development Authority ("IDMA"), and which seeks to prevent the electronic communication of false statements of facts. Under the Act, it is an offence in Singapore to say or publish a statement that one knows or has reason to believe is false or that its communication is likely to be contrary to the public interest, which interest is defined broadly, with six different possible qualifying circumstances. Although often referred to as the "fake news" law,[41] the criteria are expansive and include if the statement is likely to "incite feelings of enmity, hatred or ill-will between different groups of persons".[42] Any minister in Singapore will be empowered to issue a range of directions against any such statement where they deem it in the public interest to do so—ranging from corrections, orders to stop the communication, and/or an access blocking order. The Act also stipulates appeal mechanisms.

---

[35] http://government.ru/en/department/58/. Archived at https://perma.cc/4A9A-LP46; https://eng.rkn.gov.ru/. Archived at https://perma.cc/CQR7-DASL.

[36] See, e.g., https://www.lumendatabase.org/notices/275355. Archived at https://perma.cc/YN8A-WXWJ.

[37] https://ifade.org.tr/reports/EngelliWeb_2018_Eng.pdf. Archived at https://perma.cc/54HJ-XX54.

[38] https://www.lumendatabase.org. Archived at https://perma.cc/Y53C-H89D; Lumen is a project that collects requests to remove material from the Internet, including court orders.

[39] Center For Communication Governance, "Hate Speech Laws in India" p. 133 et. seq.. https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view; See also Report No. 267 - Law Commission of India. March 2017. http://lawcommissionofindia.nic.in/reports/Report267.pdf. Archived at https://perma.cc/GSH2-NHP6 for both a more extensive examination of current jurisprudence on hate speech in India and its effects on freedom of expression; as well as international comparisons.

[40] https://wilmap.law.stanford.edu/entries/protection-online-falsehoods-and-manipulation-act-pofma. Archived at https://perma.cc/5CFG-XE22.

[41] https://www.cnn.com/2019/10/02/asia/singapore-fake-news-internet-censorship-intl-hnk/index.html. Archived at https://perma.cc/4LXE-B6R5.

[42] Id.

On October 3, 2019, the EUCJ issued its ruling in *Eva Glawischnig-Piesczek v. Facebook Ireland Limited,* holding that "Article 15 of the EU's e-commerce directive does not prohibit EU states from ordering extremely broad injunctions against platforms like Facebook to take down offending material (and therefore allows it)." Moreover, "these injunctions can cover a wide array of material—not just Facebook posts but also reposts and "equivalent" posts—and apply worldwide."[43] Contrast this with the seemingly opposite ruling in *Google v. CNIL*, which difference may be predicated on the fact that the removal request in the Facebook case was a court order. Some commentators have also opined that the language in the *Google* v. *CNIL* case is written precisely so as to allow a result like that in *Glawischnig-Piesczek.*[44]

Pre-emptive Responsibility

Preemptive responsibility models are those that place a substantial onus on the platform itself to be responsible for removing, or in the case of notice and stay-down, for preventing not only the initial but also any subsequent posting of certain categories of content, without dictating any specific methods or terms. An excellent example of such a model is the EU Copyright Directive,[45] specifically Article 11 of that directive,[46] the text of which[47] was finalized in February 2019, adopted in late March, 2019, approved in April, and came into force on June 7, 2019. EU member states now have two years in which to pass legislation that will meet the requirements of the Directive. The Directive states that platforms, including search engines, that offer a link to a "news site" (not clearly defined) may not provide more than a "snippet" of the content available at that link; and Article 13,[48] which states, in essence, "online platforms have a duty to ensure that none of their users infringe copyright, period."[49] [NB: these articles are often referred to by their original numbers, but became Articles 15 and 17 in an update of the text.[50]]

Although there are various size thresholds a platform must meet before being subject to the EU's new obligations, with "small and medium-sized enterprises" (SMEs) exempted, the particularly stringent nature of this model, though long debated and with powerful advocates, is not something that has ever been practically implemented with respect to online content before. It is therefore not yet clear precisely how a platform would be able to comply with all of the tenets of Article 13 without creating filters that pre-screen all content prior to uploading, all while still allowing for user recourse at scale in the case of errors, as is also required. However, the Directive's authors and sponsors have repeatedly

[43] https://www.lawfareblog.com/cjeu-facebook-ruling-how-bad-it-really. Archived at https://perma.cc/AL86-2ZTH.
[44] Id.
[45] For more information about the EU Copyright Directive, see Appendix 2.
[46] https://juliareda.eu/wp-content/uploads/2019/02/Art_11_unofficial.pdf
[47] https://juliareda.eu/wp-content/uploads/2019/02/Copyright_Final_compromise.pdf
[48] https://juliareda.eu/2019/02/eu-copyright-final-text/
[49] https://www.eff.org/deeplinks/2019/03/european-copyright-directive-what-it-and-why-has-it-drawn-more-contr oversy-any. Archived at https://perma.cc/H2Y6-ME5K.
[50] https://www.cnbc.com/2019/03/28/article-13-what-eu-copyright-directive-means-for-the-internet.html. Archived at https://perma.cc/KWV2-LF74.

stated it does not require filters,[51] and in fact there are legal/expert opinions already on record[52] that a filter would be contrary to other aspects of EU law. Even some of the platforms who will likely benefit most or survive the implementation best are not supportive of the law,[53] nor are several EU countries.[54]

Also noteworthy is the "guideline" or "principles" nature of the Directive, characteristic of a takedown model that must be implemented across a range of cultures and legal regimes. Implementation is left to individual EU countries, which is itself controversial.[55]

Co-regulatory and Informal Cooperation Models

Another set of content removal models operate with substantial government involvement in the scope and targeting of content, but on a more informal basis, in the sense that removal is not mandated by law. Two models that fit this description are Google's compliance with US court orders regarding defamatory content and the United Kingdom's Counter Terrorism Internet Referral Unit (CTIRU).

In the United States, speech, whether online or not, must be found by a court to meet four distinct criteria[56] before it can be legally defamatory. Most critical or insulting online content cannot meet these tests, or has not been found to do so by a court. With that and Section 230 of the United States' Communications Decency Act in mind,[57] a user seeking to complain to Google about allegedly defamatory content either on a Google product or linked to within Google search results will be told that "Google does not remove allegedly defamatory material from our search results,"[58] but that "if you have obtained a court order which declares that the content of a web page indexed by Google is unlawful, please follow the instructions on the court order form."[59] That is, although not legally required to do so, Google will remove content a court order has found to be defamatory, presumably out of a desire for good consumer relations and wanting to "do the right thing" regarding the rule of law. [NB: Google's willingness to do this has sadly but unsurprisingly led to substantial abuse of the process.[60]]

[51] http://www.europarl.europa.eu/news/en/press-room/20190212IPR26152/agreement-reached-on-digital-copyrig ht-rules. Archived at https://perma.cc/7XA9-EAJF.
[52] https://www.ip.mpg.de/fileadmin/ipmpg/content/stellungnahmen/Answers_Article_13_2017_Hilty_Moscon-rev-18_9.pdf. Archived at https://perma.cc/NZG2-RPE6.
[53] https://www.telegraph.co.uk/technology/2018/11/12/youtube-ceo-says-new-eu-copyright-laws-threatening-creativity/. Archived at https://perma.cc/Z8FH-MZJS.
[54] https://www.parlament.gv.at/PAKT/EU/XXVI/EU/06/18/EU_61832/imfname_10895457.pdf. Archived at https://perma.cc/BTR9-NF2N.
[55] https://www.eff.org/deeplinks/2019/03/european-copyright-directive-what-it-and-why-has-it-drawn-more-contr oversy-any. Archived at https://perma.cc/H2Y6-ME5K.
[56] https://www.law.cornell.edu/wex/defamation. Archived at https://perma.cc/FDR8-YEWY.
[57] For more information regarding CDA 230, see Appendix 2.
[58] https://support.google.com/legal/troubleshooter/1114905#ts=1115655%2C1282900%2C1115974. Archived at https://perma.cc/KLK2-XZY6.
[59] Id.
[60] https://reason.com/2019/07/25/cbs-news-story-on-forged-court-orders-aimed-at-vanishing-google-search-results/. Archived at https://perma.cc/7KRP-MB4H.

The UK's Counter-Terrorism Internet Referral Unit (CTIRU)[61] organization's "work consists of filing notifications of terrorist-related content to platforms,[62] for them to consider removals."[63] However, this is not a N&TD model, since any compliance with CTIRU's notices is in theory voluntary, and the reports predicate their requests for removal on alleged violations of a platform's terms of service. However, pursuant to the "with some government involvement" facet of this model, some analysis points out that platforms, once notified, may incur responsibility or liability under the e-commerce directive. Additionally, the UK Home Office refers to all the removed content as "unlawful".[64] At this time, it is not clear what effect a successful Brexit would have on the level of pressure a CTIRU request could bring to bear on platforms.

Finally, another, less commonly used example of such a cooperative model and one that also has some aspects of a N&TD model, is that of the United States' National Center for Missing & Exploited Children (NCMEC) database and its associated protocols having to do with child sexual abuse imagery (CSAI),[65] which combine both voluntary and required aspects. Primarily funded by the U.S. Justice Department,[66] NCMEC offers a tipline for "members of the public and electronic service providers (ESPs) to report incidents of suspected child sexual exploitation," to which ESPs are legally required to report any content they discover on their services.[67] NCMEC also offers a database[68] against which ESPs can voluntarily check the content they host, in addition to any internal steps an ESP may be taking.[69] Additionally, NCMEC may on its own initiative send notices to ESPs if the presence of qualifying content is suspected.[70]

Self-Regulating Models Modeled on or Anchored in Legal Regimes

Another type of platform content regulation model is one that operates almost entirely in the private realm but that may have links, both conceptual and formal, to other legal requirements or regulatory

[61] https://wiki.openrightsgroup.org/wiki/Counter-Terrorism_Internet_Referral_Unit

[62] https://wiki.openrightsgroup.org/wiki/Counter-Terrorism_Internet_Referral_Unit#Social_media_sites

[63] https://www.openrightsgroup.org/blog/2019/informal-internet-censorship-the-counter-terrorism-internet-referral-unit. Archived at https://perma.cc/9D3U-839G.

[64] https://www.gov.uk/government/publications/counter-terrorism-strategy-contest-2018. Archived at https://perma.cc/T7PF-PX4L; https://wiki.openrightsgroup.org/wiki/Counter-Terrorism_Internet_Referral_Unit#cite_ref-1

[65] https://www.justice.gov/criminal-ceos/citizens-guide-us-federal-law-child-pornography. Archived at https://perma.cc/P49W-99CT; http://www.missingkids.com/theissues/sexualabuseimagery. Archived at https://perma.cc/HK8Y-6RWP.

[66] https://web.archive.org/web/20140821114112/http://www.opencongress.org/bill /hr3092-113/show

[67] http://www.missingkids.com/theissues/sexualabuseimagery. Archived at https://perma.cc/HK8Y-6RWP.

[68] http://www.missingkids.com/theissues/sexualabuseimagery; https://www.usenix.org/conference/enigma2019/presentation/bursztein. Archived at https://perma.cc/H7EH-8BTB.

[69] https://www.pcworld.com/article/2461400/how-google-handles-child-pornography-in-gmail-search.html. Archived at https://perma.cc/7YYK-NH7A.

[70] https://www.nbcphiladelphia.com/news/local/The-Child-Pornography-Clearinghouse-278678071.html. Archived at https://perma.cc/H6YZ-PKMJ.

regimes. Google's willingness to remove content based on a court order finding defamation could arguably be placed within this category, since to do so is a Google-internal decision.

However, the largest, most robust, well-known, and controversial[71] of these is ContentID,[72] the private monitoring, management, and removal system built by YouTube to manage its videos with respect to possible copyright infringement. Copyright infringement is a violation of federal law, and YouTube could, if it chose, rely solely on the provisions of the DMCA, discussed above. However, for a variety of reasons,[73] YouTube has spent, by some estimates, over $100 million developing the ContentID system,[74] which operates in parallel to the DMCA's mechanisms and in fact borrows a few of them.[75]

ContentID partners provide YouTube with a master list of their materials against which YouTube scans all uploads. In contrast to the DMCA, if a video is flagged as infringing, the relevant rightsholder has a wider range of options in response,including no action, reallocation of the video's monetization, and a takedown. Additionally, at any time, the rightsholder can rely on the DMCA. The original poster can dispute a ContentID takedown, but for any user who uploads at scale or has many videos, the equities of the process favor the rights-holders.[76]

Although considered the industry standard, and mimicked by other platforms such as Soundcloud,[77] ContentID is notoriously error-prone, with too many attention-getting false positives to list.[78]

Self-Regulation

In a purely self-regulating model, platforms monitor and remove content according to their particular internal needs and strictures. Platforms may even remove completely legal content, often dependent on "what kind of platform they want to be"[79] or the type of audience they intend to serve. Content is typically removed, modified, or left up without any pressure or involvement of government entities or reliance on a legal framework. Broadly, platform-initiated removals can be said to fall under "terms of service violations" and what type of content is treated as a violation or removable can vary not only from platform to platform but within a given platform over time or based on ownership. A recent example of the former internal variation is Facebook's ongoing attempts to successfully define "nudity,"[80] which has at times (but not now) included breastfeeding women. An example of the latter is

[71] https://www.bbc.com/news/technology-44726296. Archived at https://perma.cc/U2YM-M559.
[72] https://support.google.com/youtube/answer/2797370?hl=en. Archived at https://perma.cc/86WP-296K.
[73] https://publixphere.net/i/noc/page/OI_Case_Study_Intermediary_Liability_in_the_United_States. Archived at https://perma.cc/4NGC-36Y8.
[74] https://venturebeat.com/2018/11/07/youtube-weve-invested-100-million-in-content-id-and-paid-over-3-billion-t o-rightsholders/. Archived at https://perma.cc/S5S6-Y7QY.
[75] https://www.eff.org/issues/intellectual-property/guide-to-youtube-removals. Archived at https://perma.cc/M73Y-SYCD.
[76] Id.
[77] https://help.soundcloud.com/hc/en-us/articles/115003452067. Archived at https://perma.cc/6DKE-QW3K.
[78] https://www.bbc.com/news/technology-42580523. Archived at https://perma.cc/FTE9-CPSR.
[79] https://www.facebook.com/policies/ads/prohibited_content/weapons
[80] https://www.facebook.com/communitystandards/adult_nudity_sexual_activity. Archived at https://perma.cc/7PLE-B7WP.

Tumblr's shift away from adult content[81] after being bought by Verizon, where it had previously been seen and relied on by users as a viable place for that sort of material.

Platform self-regulation regarding online speech, and content in general, at whatever scale or level of attention or granularity, without fear of ensuing penalties or liability, is what the drafters of Section 230 of the United States' Communications Decency Act[82] envisioned that law would incentivize. CDA 230's provisions make it possible for service providers to moderate according to their best efforts or ability, as they see fit, without worrying whether their moderation is up to a specific standard. The Act's critical language provides for no liability "on account of—any action voluntarily taken in good faith to restrict access to or availability of material," and was intended to, among other things, "remove disincentives for the development and utilization of blocking and filtering technologies . . . ."[83] Far from freeing providers from any need or desire to moderate by providing a blanket immunity, the law, which exempts federal criminal and intellectual property claims, as well as electronic privacy claims, instead makes it possible for platforms to choose the level of content regulation in which they wish to engage without fear of being held liable for making the "wrong" choice.

What is Acceptable Content and Conduct on Wikipedia?

Removing illegal speech from Wikipedia is an important driving force in the development of processes and systems for finding and deleting problematic speech. Beyond legal considerations, there are also strong ethical reasons for Wikipedia to address harmful speech on its platform. On articles, harmful speech is antithetical to maintaining a high quality encyclopedia. Reducing the level of abuse among Wikipedians is also of vital operational importance. Maintaining an active community of editors while attracting new participants is essential to the survival of Wikipedia. If only those individuals with the thickest of skins continue to participate, the future of the platform is less promising.

For almost two decades, Wikipedia has been a living experiment in figuring out how humans can productively argue with one another and has witnessed and documented the myriad ways in which people start, perpetuate, escalate, defuse, put to rest, and avoid fights over ideas, language, and knowledge.

Content moderation on Wikipedia is governed by a sophisticated set of policies and guidelines developed by the community over many years, which have been translated into operational procedures for identifying harmful speech, removing it, and in some cases documenting the reasons a particular

[81] https://www.theverge.com/2018/12/3/18123752/tumblr-adult-content-porn-ban-date-explicit-changes-why-safe -mode. Archived at https://perma.cc/VXH3-A8VV.

[82] For more information about CDA 230, see Appendix 2. For a comprehensive history of CDA 230, see Kosseff, Jeff. The Twenty-Six Words That Created the Internet. Cornell University Press, 2019; Kosseff provides a comprehensive list of CDA 230 court opinions and other documents at https://www.jeffkosseff.com/resources. Archived at https://perma.cc/6DAG-VVS7.

[83] 47 U.S.C. § 230(b)(4)

editorial choice was taken. Although the terms "harmful speech" and "content moderation" are not frequently used, the processes and actions align at an abstract level with the content moderation practices other major platforms use. A large number of people familiar with the standards and rules, both substantive and procedural, make determinations about what content should be removed from the platform. The similarities end there.

The mission of Wikipedia, to collect and develop educational content, means that much of the garden-variety vandalism and harmful speech speech is removed regardless of its offensive nature if it is off topic or not making a constructive addition to the project. In the articles on the platform, "getting it right" is a sufficient constraint and rationale for removing much of the harmful speech that plagues other platforms.

Although the Wikimedia Foundation has a Terms of Use, which describes the rights and responsibilities that guide the foundation and its users, users are instructed to follow the sets of policies and guidelines curated by each of Wikimedia's individual projects.[84] The stated goals for English Wikipedia's exhaustive set of policies and guidelines are "to describe the community's agreed-upon best practices, clarify principles, resolve conflicts, and otherwise further the goal of creating a free, reliable encyclopedia."[85]

Most of Wikipedia's policies and guidelines pages are divided into categories for governing content, conduct, deletion, enforcement, legal issues, and procedures.

Content policies and guidelines define the scope and material that is suitable for the encyclopedia. The content policy pages explain how, for example, Wikipedia articles should maintain a neutral point of view, have verifiable sources, contain no original research, and maintain biographies of living persons with strict notability guidelines.[86]

Conduct policies and behavioral guidelines govern how editors should behave towards one another on the platform. The policies include such topics as civility, harassment, vandalism, personal attacks, dispute resolution, edit warring, and sock puppetry.[87] The behavioral guidelines cover topics that include: do not bite newcomers, be bold, assume good faith, how to respond to threats of harm, conflicts of interest, not gaming the system, and a long list of etiquette suggestions.

The talk page guidelines are a trove of information for users to learn how to collaborate on talk pages and remain civil while doing so.[88] The guidelines encourage users to remain positive and state that talk pages are a means to communicate about how to improve articles, not to "criticize, pick apart, or vent about the current status of an article or its subject." The guidelines include a list of unacceptable behavior as well; these behaviors includes insults, personal threats, legal threats, and posting other editors' personal details. Users who violate the policy by exhibiting these unacceptable behaviors are at risk of being blocked or banned.

---

[84] https://foundation.wikimedia.org/wiki/Terms_of_Use/en. Archived at https://perma.cc/24ZQ-RJYW.
[85] https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines. Archived at https://perma.cc/W9U6-7EAJ.
[86] Jemielniak, Dariusz. Common Knowledge?: An Ethnography of Wikipedia. Stanford University Press, 2014.
[87] ibid.
[88] https://en.wikipedia.org/wiki/Wikipedia:Talk_page_guidelines. Archived at https://perma.cc/5ASJ-CVUJ.

The “no personal attacks” policy page states that while no rule is “objective and not open to interpretation on what constitutes a personal attack as opposed to constructive discussion, some types of comments are never acceptable”[89]:

- derogatory phrases directed against another editor or group of editors based on race, sex, sexual orientation, gender identity, age, religious or political beliefs, disabilities, ethnicity, nationality, etc.
- using an editor’s affiliations as an ad hominem means of discrediting views or content. (Notably, this page includes an example of this type of attack. In most other cases across the policy pages where types of unacceptable behavior are listed, specific examples are not provided)
- linking to external attacks or harassment
- comparing editors to Nazis, communists, terrorists, dictators, or other infamous people
- accusations of inappropriate behavior by an editor without evidence to support the claim
- threats of legal action, threats of violence, threats to reveal personal info about an editor, or threats of actions that may expose editors to political, religious, or other persecution by a government, employer, or others

Wikipedia’s deletion policies govern the processes and decisions the community makes about deleting pages, revisions, and logs. The criteria for speedy deletion policy[90] specifies cases in which administrators have consensus to immediately delete pages or media without going through the formal deletion discussion procedures.[91] Criteria that make a page worthy of speedy deletion that are relevant to this study include patent nonsense, pure vandalism and blatant hoaxes, creations by banned or blocked users, and pages that “disparage, threaten, intimidate, or harass their subject or some other entity, and serve no other purpose,” also known as attack pages.

Wikipedia’s legal policies[92] cover child protection,[93] copyright violations,[94] libel,[95] discrimination,[96] and paid contribution disclosure,[97] among others. The platform has an oft-cited policy that it does not censor content that may be objectionable or offensive as long as the content does not violate another Wikipedia policy and it obeys the laws of the United States.[98] According to the page on Wikipedia’s non-

[89] https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks. Archived at https://perma.cc/G2CV-T2M2.
[90] https://en.wikipedia.org/wiki/Wikipedia:Criteria_for_speedy_deletion. Archived at https://perma.cc/9M3U-SKCH.
[91] https://en.wikipedia.org/wiki/Wikipedia:Deletion_process. Archived at https://perma.cc/5KVE-LAHC.
[92] https://en.wikipedia.org/wiki/Wikipedia:List_of_policies#Legal. Archived at https://perma.cc/68SV-9HJF.
[93] https://en.wikipedia.org/wiki/Wikipedia:Child_protection. Archived at https://perma.cc/Q3CC-BKX3.
[94] https://en.wikipedia.org/wiki/Wikipedia:Copyright_violations. Archived at https://perma.cc/9HQD-GTWG.
[95] https://en.wikipedia.org/wiki/Wikipedia:Libel. Archived at https://perma.cc/4D4S-UPWD.
[96] https://en.wikipedia.org/wiki/Wikipedia:Non-discrimination_policy. Archived at https://perma.cc/V9YQ-PSXR.
[97] https://en.wikipedia.org/wiki/Wikipedia:Paid-contribution_disclosure. Archived at https://perma.cc/YD4D-3L57.
[98] https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not#CENSOR. Archived at https://perma.cc/W8GL-7X5G.

censorship policy, ensuring that “articles and images will be acceptable to all readers, or will adhere to general social or religious norms, is incompatible with the purposes of an encyclopedia.”

Wikipedia’s guidelines occasionally describe what does not count as various types of unacceptable behavior or content. For example, the harassment policy page includes a section explaining what does not count as harassment:

“However, there is an endemic problem on Wikipedia of giving "harassment" a much broader and inaccurate meaning which encompasses, in some cases, merely editing the same page as another user. Therefore, it must be emphasized that one editor warning another for disruption or incivility is not harassment if the claims are presented civilly, made in good faith, and in an attempt to resolve a dispute instead of escalating one…Unfounded accusations of harassment may be considered a serious personal attack and dealt with accordingly.”

Interpretation and Implementation – Discovering Harmful Speech on Wikipedia

Wikipedia employs a multi-layered approach to discovering harmful speech on the platform. The first line of defense is automated tooling that evaluates and removes harmful content. The community has developed a number of tools along these lines including bots, automatically applied tags, and the Objective Revision Evaluation Service (ORES).[99] There are thousands of bots that have been developed by the community, but the most prominent in terms of content removal is ClueBot NG.

ClueBot NG, now in its second generation, has used a set of machine learning algorithms to revert millions of edits in its nine years of operation. The community has decided that ClueBot NG must have a high confidence that content is vandalism in order to revert, which means a lot of content that is easily recognized as vandalism by humans remains.[100]

Internal to the MediaWiki platform is ORES. Much like ClueBot NG, this service uses machine learning to assign scores to revisions based on whether or not they are damaging to the encyclopedia and whether or not they were made in good faith. Unlike ClueBot NG, these scores are not used to automatically remove content. Instead, they inform various interfaces that editors can use to prioritize which revisions they review.

Editors are also assisted in spotting and reacting to vandalism by the AbuseFilter extension, which provides "edit filters." Instead of machine-learned rules, edit filters are hand-coded rules that perform various actions if revisions conform to certain criteria. For example, some edit filters are designed to tag revisions as "possible vandalism" if they include profanity. Other filters warn newly registered users for

[99] https://www.mediawiki.org/wiki/ORES. Archived at https://perma.cc/R8B7-LWT7.
[100] https://en.wikipedia.org/wiki/Wikipedia:Bots/Requests_for_approval/ClueBot_NG. Archived at https://perma.cc/89ZQ-3888.

removing references or moving pages. Unlike Cluebot NG, which only reviews edits made to articles, edit filters can be applied to any type of page on Wikipedia.

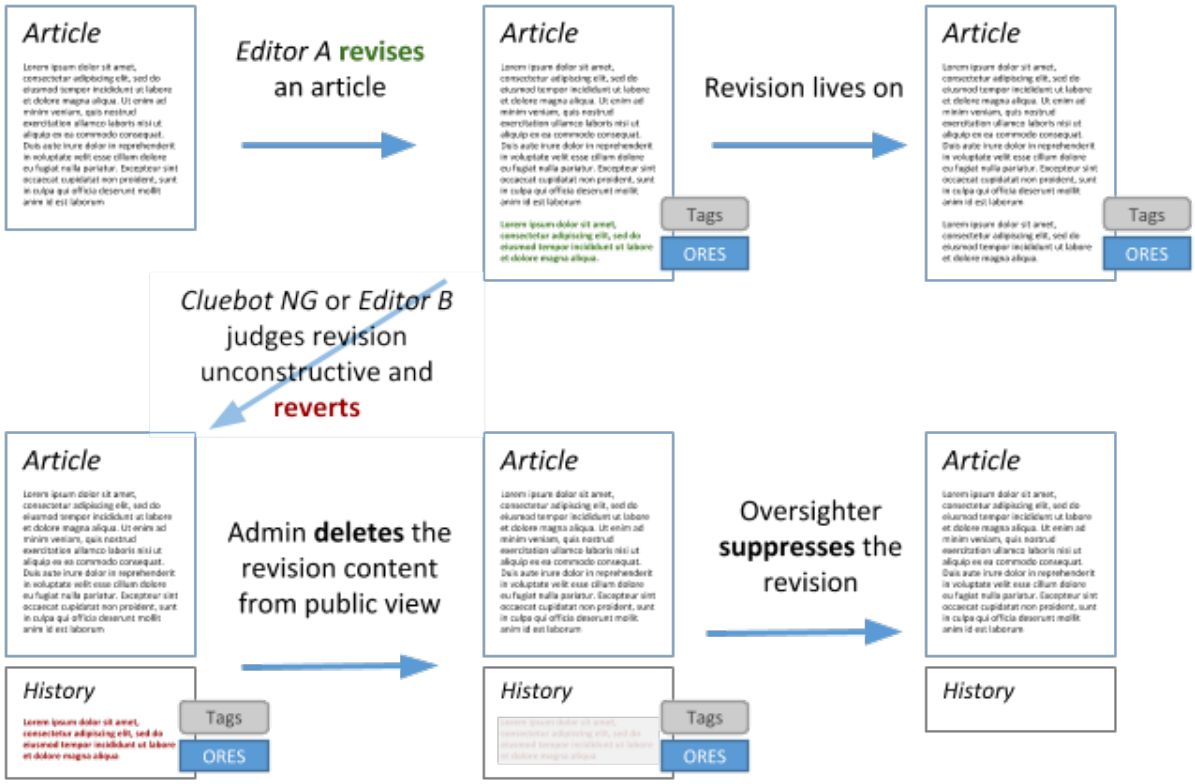


Figure 2. Article revision and content removal on Wikipedia

The second line of defense, after revisions have been removed by Cluebot NG or tagged by edit filters, are the human contributors to Wikipedia.

There are a number of different user permissions on Wikipedia that specify each user's access to advanced tools and the roles they serve on the platform. Editors are anyone who writes or modifies Wikipedia or otherwise makes contributions to the platform. Both unregistered and registered users can edit Wikipedia, but registered users are identified by their chosen account username while unregistered users are identified by their IP address. Wikipedians sometimes refer to unregistered users as "IP addresses."

Administrators (also known as admins or sysops) are editors who have been granted advanced technical permissions so that they can perform actions such as revision deletion, blocking and unblocking of registered users and IP addresses, deleting and protecting pages from editing, and granting special user

permissions.[101] [102] Oversighters are a group of 36 editors[103] with more advanced permissions that allow them to "suppress" revisions or log entries, a form of deletion that removes content from access to anyone except other oversighters. Oversighters are also able to suppress an account user name in conjunction with an account block and review all suppressed revisions or log entries.[104] CheckUsers are a group of 40 editors[105] with access to an advanced tool that allow them to determine the IP addresses of Wikipedia accounts in order to investigate and prevent cases of disruptive editing or sock puppetry.[106] [107]

There are also several specialized permissions that are automatically granted to administrators but can also be granted to veteran editors (non administrators) who submit requests.[108] For example, rollbackers are able to revert with a single click all contributions made by an editor to a single page in order to undo obvious vandalism.[109] New page patrollers review all newly created pages to identify those that do not meet the criteria for inclusion; only pages officially approved by the new page patroller group are released to be indexed by search engines.[110] A full list of roles and access levels can be found on the Wikipedia:User access levels information page.[111]

Moderator Perspectives

The following section is informed by the 16 interviews we conducted with Wikipedians with a wide range of user roles, including inexperienced editors, editors with years of experience, administrators, oversighters, and check users. The interviews were focused on gaining an understanding about how editors locate and remove harmful content, how the community as a whole interprets the policies and guidelines about how to address harmful content, and what changes, if any, could be made to improve how harmful content is addressed on Wikipedia. Appendix 1 includes a list of interview questions and extended summaries of the responses to each question.

Editors locate harmful content on Wikipedia by maintaining watch lists of articles, talk pages, and noticeboards. Editors with watchlists will receive a notification of every change that is made to each page on their watchlists. An editor will review the edits and evaluate the content according to their own interpretation of the policies and guidelines regarding harmful content. If they deem the content to be

[101] https://en.wikipedia.org/wiki/Wikipedia:Administrators. Archived at https://perma.cc/5LRC-MYQS.
[102] https://en.wikipedia.org/wiki/Wikipedia:Requests_for_permissions. Archived at https://perma.cc/7WMS-79XJ.
[103] As of November 7, 2019: https://en.wikipedia.org/wiki/Special:ListUsers/oversight. Archived at https://perma.cc/EGP2-7YZD.
[104] https://en.wikipedia.org/wiki/Wikipedia:Oversight. Archived at https://perma.cc/TJ6M-TBLG.
[105] As of November 7, 2019: https://en.wikipedia.org/wiki/Special:ListUsers/checkuser. Archived at
[106] https://en.wikipedia.org/wiki/Wikipedia:CheckUser. Archived at https://perma.cc/8MNE-LZKE.
[107] Jemielniak, Dariusz. Common Knowledge?: An Ethnography of Wikipedia. Stanford University Press, 2014.
[108] https://en.wikipedia.org/wiki/Wikipedia:Requests_for_permissions. Archived at https://perma.cc/3GXW-3C8C.
[109] https://en.wikipedia.org/wiki/Wikipedia:Rollback. Archived at https://perma.cc/XB76-BR5C.
[110] https://en.wikipedia.org/wiki/Wikipedia:New_pages_patrol. Archived at https://perma.cc/2SUF-C3RF.
[111] https://en.wikipedia.org/wiki/Wikipedia:User_access_levels. Archived at https://perma.cc/F974-3CLY.

worthy of removal, they will revert the edit. If the edit needs to be revision deleted or suppressed, they will escalate the issue as appropriate.[112]

This escalation procedure highlights the fact that content removal on Wikipedia is a matter of degree. Unlike other platforms where content is either published or not, content on Wikipedia can be limited to smaller and smaller groups of people. The first level of content removal is reversion in which the content is still accessible to the same group of people (everyone with access to Wikipedia), but its visibility is lessened significantly as it is only discoverable as part of the revision history. For some types of content this is not sufficient removal, so an administrator can "delete" a revision, which essentially limits the visibility of the content to fellow administrators. As this group still numbers more than one thousand people, oversighters, of which there are only a few dozen, can further limit the visibility to just fellow oversighters through revision suppression (or "oversight").

When editors without advanced permissions see harmful content that may need to be escalated up this chain of removal options, they have a few ways in which to notify the appropriate parties. In the course of their time on Wikipedia, veteran editors often get to know administrators who are easy to contact and are willing to respond to notices of harmful content. Editors can contact administrators directly through on-wiki email, a post to the administrator's user talk page, or Internet Relay Chat (IRC). They can also post to various noticeboards frequented by administrators such as the Administrator's noticeboard[113] and Administrators' noticeboard/Incidents.[114]

Only one participant (a non-administrator) mentioned the existence of a template to request revision deletion. However, the editor said they rarely use it because it is overly complicated. It is much easier to notify an administrator by contacting them directly through the channels noted above. According to a participant, "Approaching an admin directly is the easiest and most efficient way to get the content removed. At the end of the day, removing the content and seeing that people who post harmful content don't have a platform to do so is the goal. If I have to engage in formal mechanisms I will, otherwise informal processes work too. At the end of the day, I just want it off."

From the quantitative analysis, we see that the most common reasons cited by administrators to justify revision deletion are, in order, "Grossly insulting, degrading, or offensive material" (RD2), "Blatant violations of the copyright policy" (RD1), and "Purely disruptive material" (RD3). RD2 can be further broken into two large uses: grossly insulting material and violations of the Biographies of Living Persons policy. According to the participants, it is rare to see complaints about administrators revision deleting more potentially harmful content than necessary.

According to an oversighter we interviewed, the most common reason for suppression is copyright violation, followed by content that is grossly insulting, degrading, or offensive. Anything containing personal/private information almost always goes to oversight as well. However, the line between

[112] https://en.wikipedia.org/wiki/User:Risker/Content_moderation. Archived at https://perma.cc/9ZNK-496Q.
[113] https://en.wikipedia.org/wiki/Wikipedia:Administrators%27_noticeboard. Archived at https://perma.cc/AN5Z-QTCM.
[114] https://en.wikipedia.org/wiki/Wikipedia:Administrators%27_noticeboard/Incidents. Archived at https://perma.cc/V6SS-6GJR.

material that should be revision deleted as opposed to suppressed is somewhat unclear, according to the non-oversighters we spoke with. One interview participant said “Although the policy and criteria are usually well-followed, there aren't really any strict definitions or documented examples to go by, but there's a good sense as an admin that you know it when you see it. If it's the right thing to do, then no one will complain.”

Depending on the severity of the issue, an administrator who is working on the situation will consider whether the user should be warned or blocked. This decision depends on a number of factors, including how offensive or harmful the edits were and whether the user is likely to repeat the behavior. The administrator will examine other contributions by the same user (or IP address) and any relevant article histories for any potential follow-up actions. These potential follow up actions include additional reversions, revision deletion, suppression, page protection, IP range blocks, sock puppetry blocks, and adding articles or talk pages to watch lists. Occasionally, the editor or administrator will document the incident on talk pages or noticeboards, or they might ask other editors or administrators for advice on how to handle a particular issue.

The consensus from the participants is that if the reason why a piece of content is reverted is obvious, it is less likely to include documentation. However, the more likely it is that editors acting in good faith would disagree about whether an edit or reversion was appropriate, the more important it is to document the reasons for the change either in the edit summary, the article talk page, or both. Non-action is generally not recorded, except in circumstances when an explicit request for review is made on an administrator noticeboard. In that case, the results of the review (whether action was taken or not) will be recorded there. One participant said that there are cases where documentation of an incident can do more harm than good. “In many cases on Wikipedia, administrators act relatively autonomously, and if only one administrator thinks something needs deleting (and it can be roughly fitted within the broad policy) then they'll just do it and move on. Documentation, publicly repeating, or pointing to potentially harmful content, even if it's deleted, can sometimes be considered harmful.”

According to the participants, the frequency at which editors come across harmful content is a function of how often they go looking for it. Editors and administrators who specialize in seeking out harmful content see it often. They know exactly where to look and they have their own processes for locating it. Editors who do not seek it out do not see it as often.

How Well Does the System Remove Harmful Speech?

To evaluate the degree to which the system removes harmful speech, we draw upon the perspectives of the moderators we interviewed and content analysis. We find that the Wikipedia editor community, with help from tools that detect and remove vandalism before it reaches the public eye, does a generally effective job of detecting and removing harmful and offensive content from articles on the mainspace of Wikipedia.

Moderator Perspectives

The interview participants said that in the rare occasion that Cluebot NG or editors do not catch and remove harmful edits to articles shortly after they are made, it is possible for that content to persist for months or years. This conclusion is corroborated by the content analysis presented in the next section.

Interviewees reported that harmful content is more likely to persist on article talk and user talk pages than on mainspace articles. The potential reasons for this are multifold:

1. Cluebot NG does not operate on talk pages.
2. There are fewer editors focusing their attention on policing harmful content on talk pages than on mainspace articles.
3. It requires more effort to edit talk page conversations in a way that removes the harmful content while retaining the meaning and context of the discussion. An interview participant said that administrators will occasionally close toxic conversation threads altogether rather than attempt to remove specific passages that contain offensive text.
4. The policies and guidelines that govern harassment, personal attacks, and incivility are not interpreted and enforced consistently across the community. Unless arguments between editors use blatantly offensive language, the comments are less likely to be evaluated as a violation of the civility guidelines.
5. There is a more permissive culture of keeping up "borderline" statements on talk pages when compared to mainspace.

The participants had mixed opinions about whether the policies and guidelines provide an adequate basis for editors to address harmful content, but most agreed that they are not interpreted consistently. Policies are deliberately written and implemented to allow for individual judgement; as a result, individual editors and—to an even greater extent—administrators are given a lot of responsibility and leeway in how they decide to handle issues related to harmful content and conduct. For some this is one of the defining positive features of the platform. A participant said the policies and guidelines are not interpreted consistently "by design. The negative framing is that the guidelines are interpreted inconsistently. But the positive framing is that administrators have autonomy and discretion to handle things the way they think is best."

For other participants, the inconsistent interpretations are one of the system's biggest flaws. If a content dispute or a disagreement between editors on a talk page reaches the point where an administrator gets involved, which administrator is first to act can make a great deal of difference in the outcome of the dispute. New users are consistently judged more harshly, and veteran users are significantly less likely than new users to be blocked for adding controversial content or for incivility. One participant said "People say new editors and veteran editors should be treated the same…but not necessarily. A person who makes 100,000 appropriate edits and 10 inappropriate edits shouldn't be treated the same as someone who makes 10 appropriate and 10 inappropriate."

There was no consensus among the participants about whether there are enough editors, administrators, and oversighters to handle the amount of harmful content on Wikipedia. Several said that there are enough editors to deal with the amount of harmful content; insufficient editor activity causes problems in rote maintenance areas rather than with high profile issues. Others believe that there are not enough administrators and that burnout is an issue. The homogeneity of the editor pool is also a concern, more so than the lack of editors overall.

Several participants said that the automated detection tools like Cluebot NG do a decent job of removing obvious vandalism and harmful content so that is does not remain on an article for long. But they are skeptical that automated tools will ever be able to detect and make consistent decisions about the marginal or subtle cases of harm and incivility, considering that even the community does not have a consistent grasp on what to do about these cases.

We asked the participants to identify obstacles that the community faces in addressing harmful content and to suggest approaches or tools that would help. The participants noted the lack of diversity in the English Wikipedia community across gender, race, and other socioeconomic factors. Several suggested that the foundation and community should focus recruitment efforts to increase diversity. Future research is required to study how and to what extent the lack of diversity in the editor pool contributes to the production of harmful content or its uneven removal.

Content Analysis

Content Removal

Before an examination of harmful content removal, it is helpful to take a step back and look at the general use of content removal tools in the article, article talk, user, and user talk namespaces on English Wikipedia. This will provide a sense of scale when looking at harmful content specifically.

As we have noted above, there are three basic levels of revision removal: reversion, deletion, and suppression (oversight). Each level places greater restrictions on the number of people that are capable of viewing the revision in question. As such, we would expect each level to be used less frequently than the previous, which is consistent with what we have observed.

Removal Level	Description	No. of Users with Permission	Rate of Use
Reversion	Undoing the changes of a revision	All	1 in 30 revisions
Revision Deletion	Limiting the visibility of a reverted revision's content to administrators	1,149[115]	1 in 591 revisions
Suppression	Limiting the visibility of a reverted revision's existence to oversighters	36[116]	1 in 6,400 revisions

To estimate the volume of reversions, we looked at the use of the "undo" link that exists next to revisions, which tags the reverting revision with "mw-undo." Since early 2018 (the first instance of the "mw-undo" tag), about 1 in 30 revisions have been reverted.[117] This rate of reversion points to the presence of a stringent filter that harmful content would have to pass in order to remain. And the proportion of reverted content would certainly be higher if we were to consider all the other ways the content associated with a revision may be removed from a page.

Reversion constitutes one form of content removal as the content is removed from the article. The content is still available to anyone that cares to look for it in the edit history of the article. Revision deletion is a much stronger form of removal as the content is then only visible to administrators. Revisions are deleted at a rate of about 1 in 591.[118] These deletions were performed by 910 unique administrator accounts out of a total 2,209 administrators that have ever had the power, or about 41%.[119] Compared to reversion, those administrators must clear a higher bar when justifying a revision deletion, including citing the applicable revision deletion policy. Of the seven revision deletion codes an administrator may cite, the most cited is for "Grossly insulting, degrading, or offensive material" (RD2),

[115] The number of administrators as of November 7, 2019. https://en.wikipedia.org/w/index.php?title=Special:ListUsers/sysop&limit=2000. Archived at https://perma.cc/4HR6-2K3Q.

[116] The number of oversighters as of November 7, 2019. https://en.wikipedia.org/wiki/Special:ListUsers/oversight. Archived at https://perma.cc/EGP2-7YZD.

[117] 2.45M "mw-undo" from https://en.wikipedia.org/wiki/Special:Tags fetched on June 25, 2019. https://perma.cc/D2NV-QC8W. 72.7M total revisions from https://quarry.wmflabs.org/query/37191. Archived at https://perma.cc/B22B-93F3. First instance of "mw-undo" was Jan. 4, 2018 per https://quarry.wmflabs.org/query/37200. Archived at https://perma.cc/7NFY-8TY2.

[118] 914K revision deletions from https://quarry.wmflabs.org/query/37194. Archived at https://perma.cc/XR77-T7P6. 540.6M revisions since first deleted revision from https://quarry.wmflabs.org/query/37910. Archived at https://perma.cc/SDV8-KYFK.

[119] 914 RevDel editors from https://quarry.wmflabs.org/query/37199. Archived at https://perma.cc/XW6D-RZ44. 1,149 current administrators from https://en.wikipedia.org/wiki/Wikipedia:List_of_administrators Archived at https://perma.cc/UH3B-PUCE. Plus 1,060 former administrators from https://en.wikipedia.org/wiki/Wikipedia:Former_administrators/full. Archived at https://perma.cc/2HG3-DXZL.

followed closely by "Blatant violations of the copyright policy" (RD1), and then "Purely disruptive material" (RD3).[120] [121] RD4 indicates suppressible content which is described in more detail below, and RD5 and RD6 relate to less contentious deletion decisions. The definition of RD2 is the most closely related to definitions of harmful speech, so on the surface it looks like the plurality of the content getting revision deleted may be considered harmful.[122]

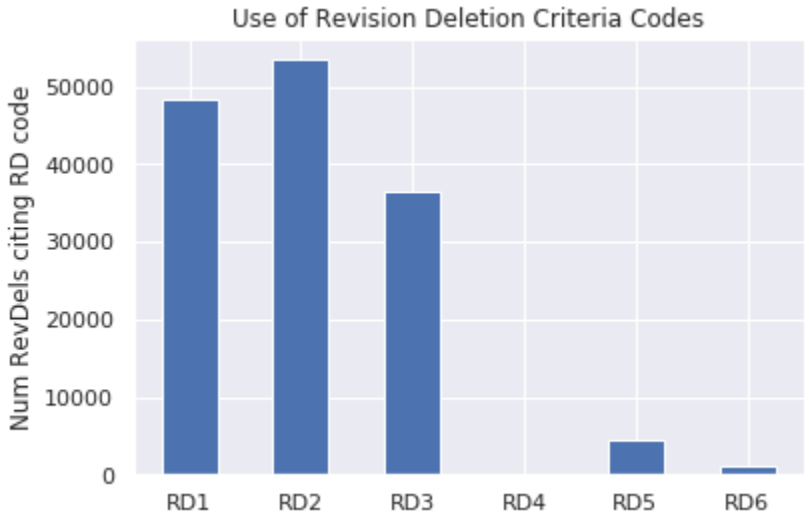


Figure 3. Use of revision deletion criteria codes

Suppression (or "oversight") is the highest level of content removal and its use is currently limited to 36 oversighter accounts.[123] Over the last twelve months, suppression use has averaged about 1,500 events per month.[124] There are potentially many revisions that are part of a single suppression event, so it is difficult to assess exactly how much content is suppressed from public data. However, our best estimate is about 1 in 6,400 revisions get suppressed.[125]

[120] The revision deletion criteria can be found at https://en.wikipedia.org/wiki/Wikipedia:Revision_deletion#Criteria_for_redaction. Archived at https://perma.cc/6C5F-TSSC.

[121] RD code usage from https://quarry.wmflabs.org/query/35439. Archived at https://perma.cc/9N53-P43Y.

[122] Further parsing of revision deletion justifications indicates RD2 citation is actually split about 70-30 between grossly insulting content and serious violations of the Biographies of Living Persons policy, but both fall more cleanly under the rubric of harmful content than other codes. See https://quarry.wmflabs.org/query/35376. Archived at https://perma.cc/R7EE-YMW3.

[123] As of November 7, 2019: https://en.wikipedia.org/wiki/Special:ListUsers/oversight. Archived at https://perma.cc/EGP2-7YZD.

[124] Statistics collected from page history of https://en.wikipedia.org/wiki/Wikipedia:Arbitration_Committee/Audit/Statistics#Rolling_six_month_Suppression_statistics. Archived at https://perma.cc/G9FB-GM2J.

[125] Count of all revisions with `rev_deleted` >8 from https://quarry.wmflabs.org/query/37194. Archived at https://perma.cc/KF82-VYD9.

It is worth noting that user-contributed content can show up in multiple ways on English Wikipedia. Each revision shows a username, an edit summary, and the actual text of the revision. Each of these pieces of content can be independently revision deleted or suppressed. Figure 4 shows the relative frequency of each of these field deletions and suppressions (on a log scale).[126]

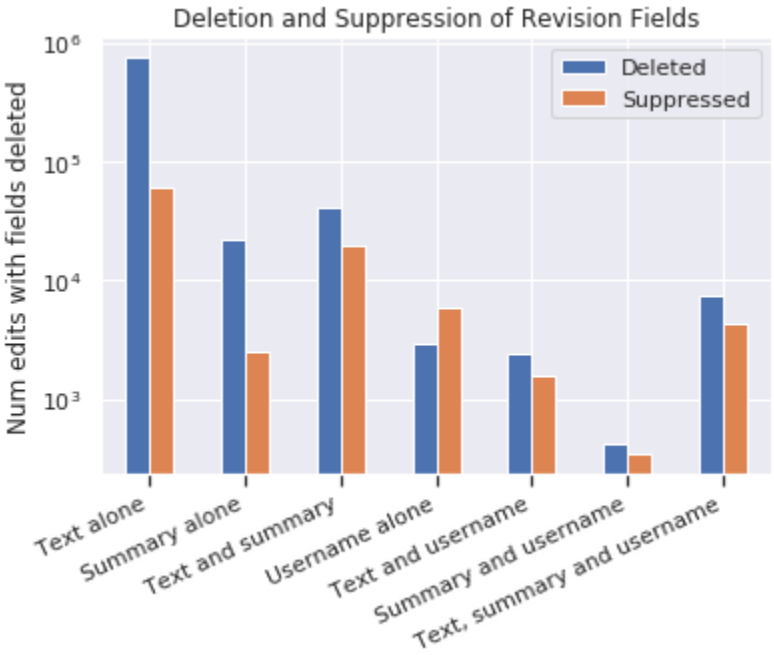


Figure 4. Revision deletion and suppression of revision fields

Harmful Content

At a base rate of 1 in 30 revisions being reverted, a lot of content is removed from English Wikipedia, but how much of the content that is removed is harmful, and perhaps more importantly, how much harmful content is missed? In attempting to locate harmful content on a large scale without tremendous resources, this study relies on a single, imperfect tool for picking out harmful content: Jigsaw's Perspective API.[127] This tool currently contains 16 models trained to predict various types of text content that may be undesirable.[128] We picked out three of these models, severe toxicity, threats, and identity-based attacks, and set fairly high thresholds for what constitutes harmful content.[129]

We started by estimating how much harmful content is on English Wikipedia in the article, article talk, user, and user talk namespaces. To accomplish this, we randomly sampled 100,000 non-deleted revisions from each of these namespaces and ran them through the process outlined in Appendix 3.

[126] From https://quarry.wmflabs.org/query/37194. Archived at https://perma.cc/242J-54LN.
[127] https://www.perspectiveapi.com/. Archived at https://perma.cc/R2YD-JTRS.
[128] https://github.com/conversationai/perspectiveapi/blob/master/api_reference.md#models.
[129] More details are in Appendix 3.

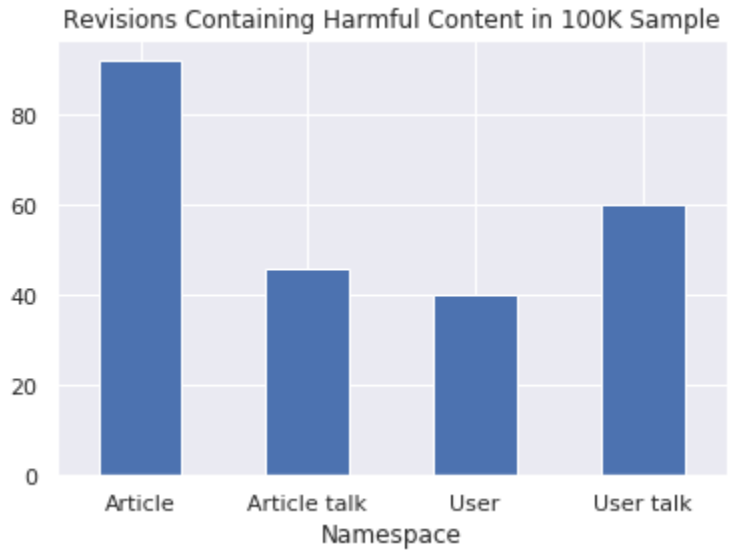


Figure 5. Revisions containing harmful speech across namespaces

As seen in Figure 5, the article namespace contains the most harmful content (92 revisions), followed by the user talk namespace (60 revisions), the article talk namespace (46 revisions), and then the user namespace (40 revisions). As a share of all content, harmful content appears in 1 of every 1,100 revisions in the article namespace and 1 in 1,700 revisions across all four namespaces.

Lifetime of Harmful Content

The revisions surfaced above were randomly sampled from all non-deleted revisions, not just revisions that contributed content to the page as it existed at the time. That means the harmful content contained in the revisions may have already been reverted (though not revision deleted or suppressed). That leads to the question of how long the harmful content that was part of these revisions existed as part of their respective pages before being reverted. To determine this, for each revision that contained harmful content, we walked through subsequent revisions until the harmful content no longer existed. Again, we did this in each of the four main namespaces. In the article namespace, none of the 92 pieces of harmful content still existed on the page of the article. The median time to takedown was 61 seconds with a maximum of 3.76 days. (See figure 6.)

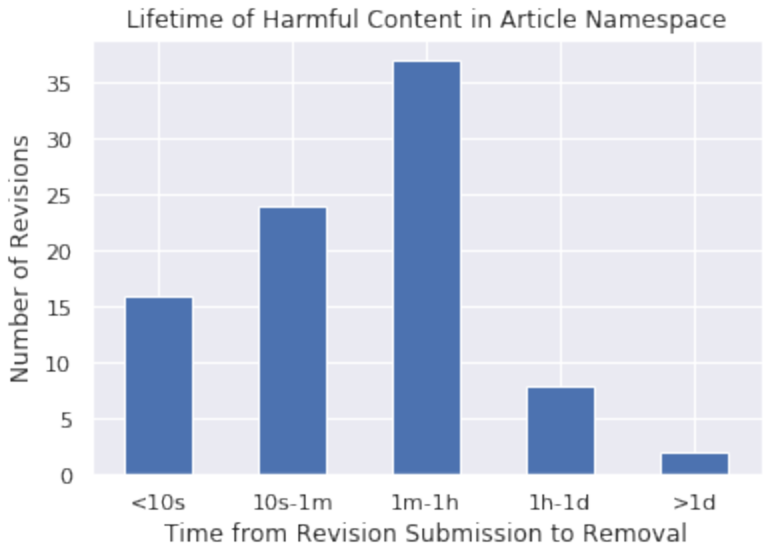


Figure 6. Lifetime of harmful content on articles

While the article talk namespace had fewer pieces of harmful content with 46, content tended to live longer (figure 7). As of this writing, one still remained up and one remained up as an edit summary. The median time to takedown was 5.3 minutes with a maximum of 268 days (excluding the one that was not removed).

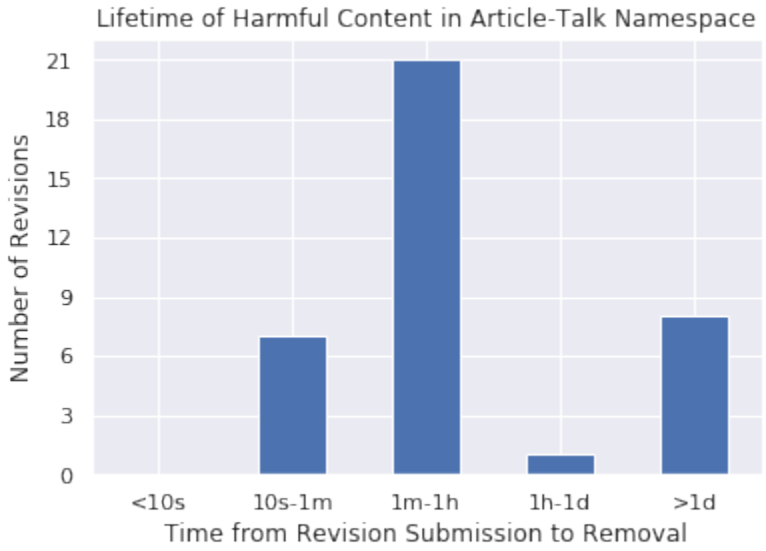


Figure 7. Lifetime of harmful content on article-talk namespace

The user namespace had the largest share of remaining harmful content of the namespaces we looked at with three out of 40 still part of the user page, but content that was taken down tended to come down quickly - the median time to takedown was 59 seconds with a maximum of 90.2 days (excluding the three that have not been removed). (See figure 8.)

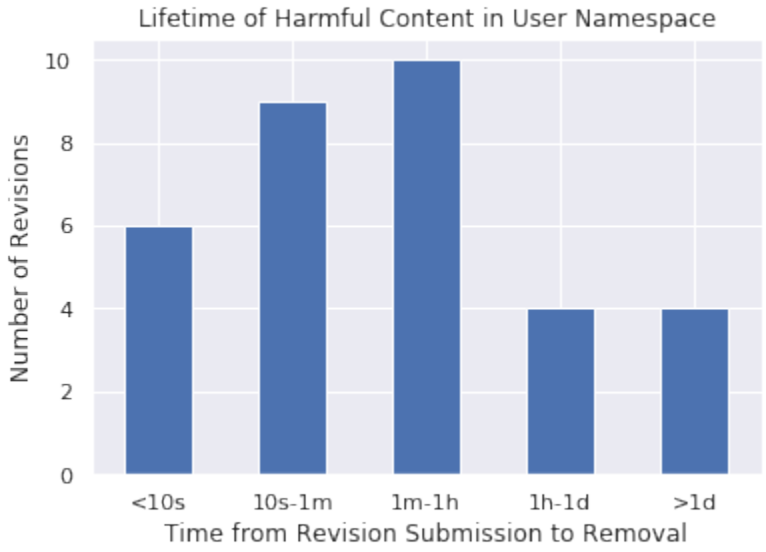


Figure 8. Lifetime of harmful content in user namespace

Finally, in the user talk namespace (figure 9), of the 60 instances we identified, four remained up on the user's talk page. The median time to takedown was 73 seconds with a maximum of 23.1 days (excluding the four that were not removed). On its face, these numbers appear to contradict the finding from "Ex Machina: Personal Attacks Seen at Scale" by Wulczyn et al. that only 18% of personal attacks in the user talk namespace result in moderation. This difference can be explained by a number of factors. The biggest difference between these two studies is that Wulczyn et al. define moderation as the warning or blocking of the attacking user. We are instead looking at the removal of the content. Our bar for what constitutes harmful content is also much higher than the bar for a personal attack. Finally, they removed bot and administrative revisions, which constituted between 20-50% of all revisions depending on namespace. We did not.

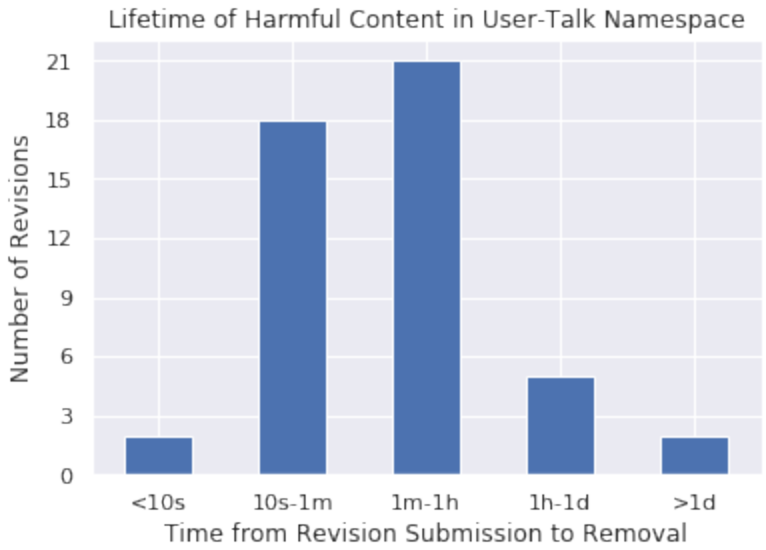


Figure 9. Lifetime of harmful content in user-talk namespace

These results are summarized in table 1, below.

	N	Number	Proportion	Non-reverted	Median time to removal	Max time to removal
Article	100,000	92	1 in 1,100	0	61s	3.76 days
Article talk	100,000	46	1 in 2,200	1	319s	268 days
User	100,000	40	1 in 2,500	3	59s	90.2 days
User talk	100,000	60	1 in 1,700	4	73s	23.1 days

Table 1. Revisions with harmful speech across namespaces

Across all four namespaces, 230 of 238 revisions containing harmful content were reverted by the time of writing, or 96.6%. The median time to removal from the main page was 75 seconds.

Harmfulness of Deleted Revisions

It is possible that the volume of harmful content introduced to English Wikipedia is higher than our estimates above because harmful revisions may have already been hidden from public view by deletion or suppression. To test this theory, and to test how our method of detecting harmful content relates to the use of revision deletion criteria codes, we gathered revisions in real-time from April through July 2019. We then followed up toward the end of the study to look at the content that had been revision deleted. Over the study period, we collected more than 9 million revisions of the 12.6 million that took place.[130] 6,700 revision deletions took place during the same period, which amounts to about 1 in 2,100 though that might be inflated by the fact that revisions that were deleted may have been created prior to the beginning of our data collection.[131] Because we did not collect all revisions that took place in our period of study, we were able to analyze content for approximately 3,000 of the 6,700 deleted revisions. Of those, 54 were classified as harmful content, or about 1 in 56. This is much higher than the 1 in 1,700 we saw across all four namespaces, suggesting harmful content is revision deleted at a higher rate than average content. If we use 1 in 1,700 as an estimate for the proportion of revisions including harmful content, we would expect about 7,400 of the 12.6 million revisions to include harmful content. If only 54 of these were revision deleted, it is unlikely that revision deletion is hiding a significant portion of harmful content from our analysis.

We are also able to assess how the use of revision deletion criteria codes relate to our method of detecting harmful content. Of the 414 revisions deleted using RD2 as justification, 39 were classified as harmful (9%). 14 of 154 revision deletions tagged as RD3 ("purely disruptive") were classified as harmful (9%). None of the 2,397 revision deletions using the RD1 code ("copyright infringement") were classified as harmful speech.[132] As one might expect, content tagged by editors as copyright infringement was not classified as harmful. The distinction between "grossly insulting" and "purely disruptive" content is less clear. Both categories were classified as harmful in approximately equal proportions.

The Editors who Delete Harmful Content

Who reverts the harmful content we identified? In the article namespace, we saw that 83 unique editors reverted the 92 pieces of harmful content we had identified. Only three of those editors reverted more than a single piece of content. ClueBot NG reverted 8, ClueBot reverted 2, and VoABot reverted 2. This same pattern holds across the other three namespaces; a very small number of editors revert multiple pieces of harmful content, while the vast majority is distributed across a wide number of editors. In fact, in the other namespaces, no single editor reverted more than two, which we interpret as a testament to the effectiveness of ClueBot NG in the article namespace, and which also highlights its absence in other namespaces.

[130] https://quarry.wmflabs.org/query/38028. Archived at https://perma.cc/G88L-AA8G.

[131] https://quarry.wmflabs.org/query/38027. Archived at https://perma.cc/LQ4T-EMEF.

[132] The ratios of use of the codes differs substantially from the collected stats included previously. The reason for this is not clear.

Where Harmful Content Might Concentrate

One question we were interested in was: How much harmful content is not getting removed? We took one pass at that above, but it is worth reiterating that identifying harmful content is a very difficult problem, even at smaller scales. If harmful content were easy to define and identify, there would be none on the large platforms that deem it undesirable. In an effort to narrow down places to look for harmful content, we asked a number of people involved with English Wikipedia where they thought there might be a higher prevalence. The responses include:

1. Revisions on pages in the "Objectionable Content" category[133]
2. Revisions on pages in the "Controversial Topics" category[134]
3. Revisions marked as "possible vandalism" and the like by tag filters[135]
4. Revisions by editors or on pages referred to on the Administrator's Noticeboard for Incidents[136]
5. Revisions reverted by ClueBot NG[137]
6. Revisions almost reverted by ClueBot NG[138]
7. Revisions by editors who have recently been reverted
8. Revisions by editors who have had revisions deleted in the past
9. Revisions with poor ORES scores, especially on the "bad faith" metric
10. Revisions from "breaking news" articles[139]
11. Revisions from pages with editing protection prior to protection
12. Revisions from the user and user-talk pages of administrators
13. Combinations of the above (e.g. revisions on talk pages of administrators by users who have recently been reverted)

---

[133] https://en.wikipedia.org/wiki/Category:Wikipedia_objectionable_content. Archived at https://perma.cc/6X5G-KWN4.
[134] https://en.wikipedia.org/wiki/Category:Wikipedia_controversial_topics. Archived at https://perma.cc/Z2LQ-KHQH.
[135] https://en.wikipedia.org/wiki/Special:Tags. Archived at https://perma.cc/D2NV-QC8W.
[136] https://en.wikipedia.org/wiki/Wikipedia:Administrators%27_noticeboard/Incidents. Archived at https://perma.cc/V6SS-6GJR.
[137] https://en.wikipedia.org/wiki/Special:Contributions/ClueBot_NG. Archived at https://perma.cc/SP33-RU2F.
[138] https://en.wikipedia.org/wiki/User:ClueBot_NG#ClueBot_NG_IRC_Feeds. Archived at https://perma.cc/VKG6-R29L.
[139] https://en.wikipedia.org/wiki/Portal:Current_events/Sidebar. Archived at https://perma.cc/HWJ6-G4KX.

The presence or absence of harmful content in each of these spaces may point to policy prescriptions or other possible community interventions, but collecting revisions from each proved to be a time consuming process. Instead, we chose to sample 10,000 revisions from the first three spaces and include the remaining as areas for future research. "Objectionable content" pages had 43 harmful revisions (1 in 232), "controversial topics" pages had 8 harmful revisions (1 in 1250), and revisions with vandalism tags had 265 (1 in 38).[140] The gap in the amount of harmful content between "objectionable content" and "controversial topics" is somewhat surprising given their similar constructions (automatically curated lists of pages that include certain boilerplate wiki text, e.g. "Wikipedia is not censored...") The high prevalence of harmful content in revisions tagged as vandalism suggests that the tag filters applying these tags are effective in identifying harmful content.

Since revisions with vandalism tags had the highest proportion of harmful content of spaces we looked at, we decided to look at revisions within this space that our process did not flag as harmful to get a sense of how much harmful content our detection process may have missed. We manually reviewed a random sample of 100 revisions of the 9,734 revisions and found 22 that could qualify as harmful content under our definition. Specifically, 20 were defamation and two were identity-based hate. Because defamation is not easily identified by the models we used, this was a known limitation. This simple exercise suggests our process missed about 2% of what we would expect it to pick up and about 22% of what it would pick up in the ideal case. The fact that these revisions were flagged by editor-coded tag filters and missed by our process due to technical limitations speaks to the power of Wikipedia's more developed, community-based system.

Determining how well English Wikipedia is doing its job of removing harmful content from the platform is a tricky question to answer. Locating harmful content is a required part of coming to that answer, but it is also something the community of English Wikipedia has been working at for many years. If the answer was "worse than state-of-the-art algorithms can do," our analysis could more meaningfully draw a line between successful and unsuccessful identification and removal of harmful content. Fortunately, that is largely not the case. We can say that a very high proportion of harmful content that is detectable by algorithms is removed and removed quickly.

Media Coverage of Content Moderation

Another perspective on content moderation on Wikipedia is the proportion of media coverage of the platform that is devoted to content moderation issues compared to other platforms. In figure 10, we show results of a query run on the Media Cloud platform.[141] For each platform in the graph, we searched for all articles in top U.S. media outlets over the last four years whose titles contained the name of the platform. Within each resulting set of articles (one for each platform), we determined what

[140] Objectionable content revisions sampled from https://quarry.wmflabs.org/query/37035. Archived at https://perma.cc/K2F6-GG4T. Controversial topics revisions sampled from https://quarry.wmflabs.org/query/37041. Archived at https://perma.cc/6QLC-NT3K. Vandalism-tagged revisions sampled from https://quarry.wmflabs.org/query/37044. Archived at https://perma.cc/Y8MS-EVD4.
[141] https://mediacloud.org. Archived at https://perma.cc/RF7Z-YJ3B.

percentage contained one or more of the phrases "hate speech," "harmful speech," "harmful content," and "content moderation." This perspective suggests that content moderation on Wikipedia is less controversial than that of other platforms, commanding a lower proportion of media attention than Facebook, Twitter, YouTube, and Google.

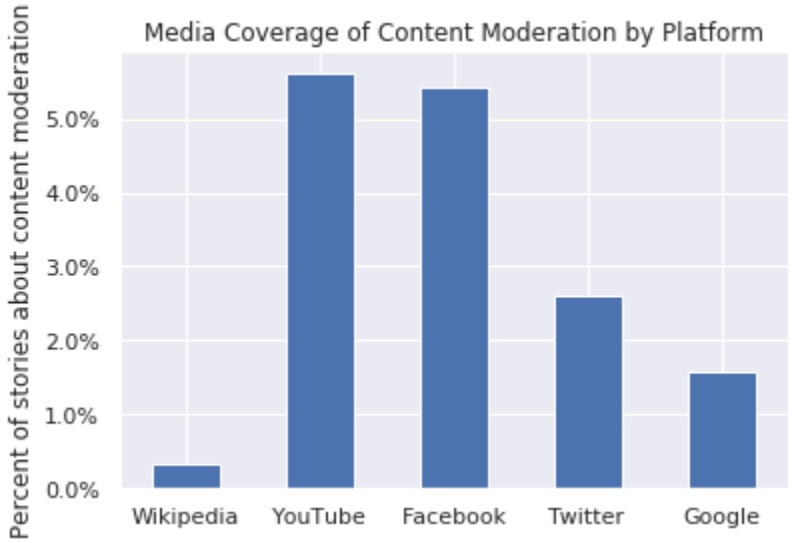


Figure 10. Media coverage of content moderation by platform

Limitations

This study is based on both qualitative and quantitative methods and both have limitations that should inform the interpretation of the results and point toward future refinements.

Due to the many software changes that have taken place over Wikipedia's long existence, it can be difficult to know when certain features were added or removed and when and how those changes manifest in the database. For example, the number of deleted revisions looks very different depending on the table one uses, either "logging" or "revision".

As outlined in Appendix 3, we used three of 16 possible classifiers. Those three classifiers together cover only part of our larger definition of "harmful content" outlined above. As such, our results only apply to these types of content, not our full ontology of harmful content. When using these classifiers, we set high thresholds for confidence, so the content that is classified as harmful tends to be particularly clear cut or extreme.

It should also be noted that these models are state-of-the-art in terms of performance, but they are not as good as humans at picking out the harmful text in arbitrary corpora. There is a growing body of research documenting the weaknesses of algorithmic detection of harmful speech in general and those

of the Perspective API specifically.[142] Of particular concern is the ability of algorithms to propagate biases present in the data used to train them.[143] [144] [145] To help mitigate the risks posed by these weaknesses, we reviewed all comments flagged as harmful for false positives as well as estimated the false negative rate. We observed significant false positives and negatives, especially in the threat model.[146] More details are provided in Appendix 3. While these weaknesses and biases constrain our analysis, we are more tightly constrained by the more general sub-human classification performance. By relying on code with sub-human performance, we are not able to assess how well English Wikipedia removes harmful content in the general case. We cannot tell how much better English Wikipedia is doing than the algorithm, so the best we can do is assess whether or not the platform meets the baseline set by state-of-the-art algorithmic content moderation.

The findings from the interviews are based primarily on feedback from the first 16 editors who volunteered to be interviewed. Although the participants who volunteered encompassed a broad spectrum of user roles, including relatively inexperienced editors, editors with years of experience, administrators, oversighters, and check users, we suspect that there was a degree of self-selection bias as a result of the recruitment methods. Therefore, we do not know how closely the opinions of the participants represent the Wikipedia community as a whole. Similar studies in the future should strive to recruit a deliberately diverse range of participants that includes more women, minorities, and editors of English Wikipedia from around the world.

Despite the limitations of our current methods, the results of each generally line up, which can give us more confidence in our results than either method could alone.

Summary and Discussion

We conclude from this study that Wikipedia is highly effective at quickly identifying and removing harmful content from articles on the platform despite the challenges of scale and openness of the site. The automated tools complement the vigilance and commitment of a large number of volunteer editors that remove harmful speech that vandals and trolls attempt to introduce. The decentralized structure empowers editors to act quickly and decisively to interpret what constitutes damage and harm on the encyclopedia. Decentralization also results in inconsistency in the application of the guidelines and standards that guide moderation decisions, although this appears to be more than offset by the benefits and agility of autonomous context-guided decision making. The general consensus among those we

[142] Hosseini, H., Kannan, S., Zhang, B., Poovendran R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. https://arxiv.org/abs/1702.08138. Archived at https://perma.cc/2EBE-AXGS.

[143] Binns R., Veale M., Van Kleek M., Shadbolt N. (2017) Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In: Ciampaglia G., Mashhadi A., Yasseri T. (eds) Social Informatics. SocInfo 2017. Lecture Notes in Computer Science, vol 10540. Springer, Cham

[144] https://twitter.com/jessamyn/status/900867154412699649. Archived at https://perma.cc/RTQ4-B2VK.

[145] Sap, M., Card, D., Gabriel, S., Choi, Y., Smith N. (2019). The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1668–1678.

[146] This is noted by others as well: https://blog.wikimedia.org/2017/02/07/scaling-understanding-of-harassment/. Archived at https://perma.cc/WW8P-GSNW.

interviewed is that the inconsistency in application is a small price to pay and that enforcing more clearly articulated rules for acceptable content would be costly and counter-productive. Facing a trade-off that Robyn Caplan frames as consistency versus context, content moderation of articles on Wikipedia fits squarely into the context-focused approach.

The successes of removing harmful speech on Wikipedia are tied closely to the tools and processes designed to fight off any manner of vandalism. These successes should not allow us to lose sight of how frequently contentious the process is and the extent of ongoing debates over policy and process. These conclusions also do not speak to the heavy burden placed on volunteer editors and the reliance of the system on the most productive and committed editors.

The more difficult and ongoing challenge for Wikipedia is moderating and addressing harms in the interactions of the community itself and malicious interlopers attacking Wikipedians rather than articles. The roots of this issue are entangled with the core mission and collaborative open nature of Wikipedia. The efforts to foster an amicable and civil community that is able to negotiate whose contributions remain and whose are removed is every bit as formidable as collaborating in the production of knowledge. The continued viability of Wikipedia depends on maintaining an empowered and committed community, which in turn depends on the community creating and sustaining the social norms that govern conduct and discourage harmful interactions. Wikipedia is also a platform that seems to be somewhat calcified against institutional changes, especially those proposed by newer editors.[147] As Halfaker et al. suggest, "Wikipedia has changed from the encyclopedia that anyone can edit to the encyclopedia that anyone who understands the norms, socializes himself or herself, dodges the impersonal wall of semi-automated rejection, and still wants to voluntarily contribute his or her time and energy can edit."[148]

Still, our study shows that the most severe forms of harmful speech are relatively rare and the vast majority are quickly removed. Although we do not have direct observations of this dynamic, the difference in this case is that the damage of harmful attacks to members of the community is more likely to have a profound and lasting effect than the harms of ephemeral vandalism on articles.

Much of the discordant interactions on Wikipedia fall below the threshold of egregious and obvious forms of harmful speech. We are not able to assess the less overt and blatant attacks against volunteer editors using the automated tools we employed in this study. More fully understanding the depth and extent of such attacks will likely require other approaches.

The relative success of Wikipedia in policing its content stand in stark contrast to the ongoing struggles of the large commercial platforms—Facebook, Twitter, and YouTube—to monitor their platforms using a combination of automated tools, community flagging, and paid moderators. In the industrial-scale moderation models described by Tarleton Gillespie and others, content moderators are paid workers that are meant to faithfully follow specific directions on what stays up and what goes down.

[147] Jemielniak, Dariusz. Common Knowledge?: An Ethnography of Wikipedia. Stanford University Press, 2014.
[148] Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2013). 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.

Unfortunately, the community-reliant approaches utilized by Wikipedia do not translate to commercial platforms.

There are several future extensions to this work that would help us to better understand the effectiveness of content moderation on Wikipedia and identify potential mechanisms to improve the tools and community structures. One potentially fruitful area of study would be to better understand how removal efforts and account suspensions are influencing the incidence of harmful speech. This would require a more intensive data collection effort over time. The quantitative analysis in this study averaged results over the entire history of English Wikipedia. Given the substantial changes in standards, policy, tools, and implementation over that period, future research might assess the trends over time and evaluate how the current systems are performing compared to prior periods. Another interesting question that is beyond our reach in this study is quantitatively measuring how harmful speech on Wikipedia compares to other online platforms and assessing how it compares to other fora online and offline.

Appendix 1: Perspectives of Wikipedians – Interview Results

We interviewed 16 Wikipedia editors about the processes and guidelines for content revision, content deletion, and quality control of English Wikipedia. The interviews focused on gaining an understanding of the community's policies and decision making about how to handle harmful content both on articles and talk pages. The focus of the conversations were not on the actions of individual editors and administrators but rather to understand how the system operates overall.

Below are the key takeaways from the conversations. This section includes a list of interview questions, extended summaries of the responses to each question, and anonymous quotations from the interview participants.

Key Findings

- Editors and administrators use a variety of methods to locate harmful content including long watch lists of articles, talk pages, and noticeboards.
- The frequency at which editors come across harmful content is a function of how often they go looking for it. Editors and administrators who specialize in seeking out harmful content see it often. They know exactly where to look and they have their own processes for locating it. Editors who don't seek it out don't see it often.
- Harmful content is more likely to persist on talk pages than articles.
- The line between material which should be revision deleted as opposed to suppressed is somewhat unclear, according to the non-oversighters we spoke with.
- The consensus is that there is not a lot of harmful content which persists for any length of time, but it does happen rarely and it can sometimes persist for months or years.
- There is not much public disagreement on Wikipedia about content that gets revision deleted. It is rare to see complaints about the community removing too much potentially harmful content.
- The participants had mixed opinions about whether the policies and guidelines provide an adequate basis for editors to address harmful content. Several said that the policies are deliberately written and implemented to allow for individual judgement.
- Administrator actions related to removing harmful content are rarely disagreed with or undone without good reason. If a situation reaches the point where an administrator is involved, which administrator acts can make a great deal of difference. For some, this is one of the defining features of Wikipedia content moderation, and for others it is a design flaw.
- The understanding and enforcement of biographies of living persons (BLP) notability guidelines are a common cause of disagreement. Some in the community think the guidelines are applied

too strictly, while others argue that the notability guidelines are not strict enough. A few participants suggested raising the notability guidelines so that there are less BLP pages overall.

- The consensus among most of the participants is that the guidelines and policies about harmful content are not interpreted consistently across the community. In unambiguous cases of defamation or clear vandalism, the community does a consistent job. In marginal cases, especially those involving incivility between editors, there are a wide range of opinions and outcomes.

- There was no consensus among the participants about whether there are enough editors, administrators, and oversighters to handle the amount of harmful content on Wikipedia. Several said that there are enough editors to deal with the amount of harmful content; insufficient editor activity causes problems in rote maintenance areas rather than with high profile issues. Others believe that there are not enough administrators and that burnout is an issue. The homogeneity of the editor pool is also a concern, more so than the lack of editors overall.

- Several participants said that the automated detection tools like Cluebot NG do a decent job of removing obvious vandalism and harmful content before it reaches public view. But they are skeptical that automated tools will ever be able to detect and make consistent decisions about the marginal or subtle cases of harm and incivility, considering that the community doesn't even have a good grasp on what to do about these cases.

- The diversity problem of English Wikipedia goes well beyond the gender gap. Participants suggested that the foundation and community should focus recruitment efforts on addressing diversity with many other socioeconomic factors in mind too.

- Sock puppetry remains a big problem. Two editors suggested that eliminating the ability for non-logged in users to edit could be a solution but acknowledged how controversial such a change would be.

- Veteran editors have a low tolerance for harmful edits that are made by users that have not logged in (identified only by the user's IP address) and new users. Established users are significantly less likely than new users to be blocked for adding controversial content or for incivility.

Question 1: How do editors or administrators locate harmful content that should be removed from Wikipedia?

The participants noted the following locations and methods to locate harmful content.

- Watch lists, especially of known hot spots. These include articles and talk pages, although participants said that watching articles is more common than only watching the associated talk pages.

- The new pages list is monitored by the new page patrol

- Administrator noticeboards:
    - WP:AN
    - WP:ANI
    - WP:AIV
    - WP:UAA
    - WP:PFPP
- Administrators (and Wikipedians with more advanced permissions) will get private messages and messages on user talk pages from other editors and administrators about content.
- Abuse filter logs/edit filters
- One administrator said they occasionally use either Google or Wikipedia's search engine to find problem strings
- Largest volume of damaging content is reverted automatically by bots.
- OTRS tickets
- Help desks
- The impression is that most administrators have similar watchlists and there is a lot of overlap, but administrators also branch into specialized collections and niches.
- There are vandals who will target pages that are suddenly highly visible because they are featured on the main page
- Vandals will create new accounts with names that refer to recent atrocities. A participant gave the example "User:Weebs killed by a fire at kyoto animation"
- Other pages that are visited to find harmful content:
    - Special:AbuseLog
    - Special:Log/newusers
    - Special:NewPagesFeed
    - Category:Candidates for speedy deletion
    - Special:Log/spamblacklist
    - Wikipedia:Sockpuppet investigations
    - Special:RecentChanges

Key Quotes

“I also have a large collection of user (and user talk) pages on my watchlist. These are both subject to vandalism (typically classifiable as abuse, threats, or harassment), and contain discussions about where to find problems, including people asking for help. Administrator talk pages are always especially interesting, and they also get a lot of abuse (like death threats and stuff).”

“The spam blacklist is not only used for spam. If someone is trying to add pornhub links to biographies (which would be prevented and logged by the blacklist), then you can guarantee they'll be trying some other harm.”

Question 2: How frequently do you encounter content that should potentially be removed from Wikipedia because it is harmful?

The answers varied widely across the participants depending on their role, area of interest, and how often they actively look for harmful content.

The consensus is that editors and administrators who specialize in seeking out harmful content see it often. They know exactly where to look and they have their own processes for locating it. Editors who don’t seek it out don’t see it that often. The bots are widely viewed as being fairly good at locating and removing obvious vandalism and offensive content. Editors who hang around highly trafficked areas might see signs that harmful content has been removed, but issues are dealt with quickly because of how many eyes are constantly on these areas.

Question 2a: Do you more commonly find harmful content on articles or talk pages?

The answers to this follow up questions also varied widely depending on role and editor focus. Several participants noted that harmful content on talk pages will often go unnoticed for longer because it is less visible and requires more effort to edit in a way that removes the harmful content while retaining the meaning of the conversation. Editors are often uncivil in edit summaries as well.

Key Quotes

“Talk pages tend to be slightly more problematic, since it is there that arguments between “sides” can become heated. The articles are watched more assiduously.”

“In general, I’m less concerned with personal attacks or near-personal attacks then I am with POV skewing of content or the insertion of false information into articles, which I believe to currently have the most potential to harm Wikipedia.”

“The vast majority is in the articles, maybe like 97% plus. We do get some in the talk pages, and it will sometimes go unnoticed for longer. Talk page vandalism is a lot less visible, and requires more effort to make meaningful. There's quite a lot of harmful content that users get on their user talk pages.”

“More so on talk pages. There’s a lot of argument. Editors who try to put harmful content on the platform got savvy to the fact that they are going to be immediately reverted if they try to do it on an article. But there’s a more permissive culture of keeping up borderline statements on talk pages; they can push back against more effectively against removal until an administrator gets involved.

Question 3: What are the steps editors (or administrators) take when content is located that might need to be removed?

The participants mentioned the following steps to address harmful content.

1. Evaluate the content.
2. Revert the content as appropriate.
3. Consider whether the user should be blocked or warned. This can depend on many things including how offensive the edits were and whether more are likely.
4. Consider if edits need revision deletion or oversight and escalate as appropriate.
5. Examine other contributions by the same user (or IP range).
6. Examine any relevant article histories for any further problems or potential action. For example, more removals, deletion, revision deletion, page protection, range blocks, sockpuppetry blocks, watchlisting.
7. (optional) Follow up or make any related reports. In some cases, though it is not very common, it can be useful to ask other editors/administrators for "more eyes" on a situation.
8. (optional) Work on any affected articles to see if future occurrences can be prevented.

Some editors and administrators will ask editors to “self-revert” if the offending comment is on a talk page in order to give the offending editor enough rope to save themselves; if the offending editor will not revert their comment, then the matter can be escalated to the Administrators’ noticeboard/Incidents.

Question 4: What is the criteria and process for deciding whether a piece of content should be flagged for revision deletion[149] or oversight[150] rather than simply reverting or editing the content?

There are several ways in which administrators are notified that a piece of content needs to be reviewed for potential revision deletion. Many veteran editors know administrators who are easy to contact and are willing to respond to a notice of harmful content. Editors can contact administrators directly through on-wiki email, a post to their user talk page, or Internet Relay Chat (IRC).

Only one participant (non-administrator) mentioned the existence of a template to request revision deletion. However, the editor said they do not use it because it is overly complicated. It is much easier to notify an administrator by contacting them directly through the channels noted above.

The line between material which should be revision deleted as opposed to suppressed (oversighted) is somewhat unclear, according to the non-oversighters we spoke with. Even one administrator said they never use revision deletion. They either revert an edit, or if it is particularly offensive, they will send it to oversight. The most common reason for oversight is copyright violation. The second most common reason is that the content is grossly offensive, degrading, or insulting. Anything containing personal/private information almost always goes to oversight as well.

Oversighters have a private queue in the Open Ticket Request System (OTRS) through which oversighters are notified of edits that need to be suppressed. Oversighters also use a conversational mailing list to ensure consistent interpretation of policies. The standard for oversighters is to "suppress first, discuss later". Although it is possible to unsuppress content that was later decided should not have been suppressed, it happens very rarely. An oversighter estimated it occurs once a month out of the roughly 500 suppression events per month.

Overall, the participants gave no indication that there are any loud calls from the community to reform the revision deletion and oversight process.

Key Quotes

"Although the policy and criteria are usually well followed, there aren't really any strict definitions or documented examples to go by, but there's a good sense as an administrator that you know it when you see it. If it's the right thing to do then no one will complain. It's often said that administrators were chosen for their judgment and their knowledge of the rules, rather than their ability to follow them. It's a theme you'll find throughout Wikipedia."

[149] https://en.wikipedia.org/wiki/Wikipedia:Revision_deletion#Criteria_for_redaction. Archived at https://perma.cc/6C5F-TSSC.
[150] https://en.wikipedia.org/wiki/Wikipedia:Oversight#Policy. Archived at https://perma.cc/TJ6M-TBLG.

“My criteria for this is pretty simple: offensiveness. I have a pretty high bar for not being upset about cursing or petty name calling, so if I think it’s offensive, I can be fairly sure that others will think so as well.”

“Approaching an administrator directly is the easiest and most efficient way to get the content removed. At the end of the day, removing the content and seeing that people who post harmful content don’t have a platform to do so is the goal. If I have to engage in formal mechanisms I will, otherwise informal processes work too. At the end of the day, I just want it off.”

Question 5: How are decisions about removing or not removing a piece of harmful content documented?

The consensus from the participants is that if the reason why a piece of content is reverted is obvious, it is less likely to include documentation. However, the more likely it is that editors acting in good faith would disagree about whether an edit or reversion was appropriate, the more important it is to document the reasons for the change either in the edit summary, the article talk page, or both. Non-action is generally not recorded, except in circumstances when an explicit request for review is made on an administrator noticeboard. In that case, the results of the review (whether action was taken or not) will be recorded there.

If a piece of harmful content is reverted by cluebot, the reversion log doesn’t include documentation about the reason for the reversion.

One participant added that documentation of decisions has administrative labor costs that can add up quickly. In the opinion of some Wikipedians, that labor is better spent improving content rather than documenting decisions.

Key Quotes

“In many cases on Wikipedia, administrators act relatively autonomously, and if only one administrator thinks something needs deleting (and it can be roughly fit within the broad policy) then they'll just do it and move on. Documentation, publicly repeating, or pointing to potentially harmful content, even if it's deleted, can sometimes be considered harmful.”

“There’s no log of an administrator looking at a situation and doing nothing. It leaves no trace they were there. Another administrator may come along and effectively overturn the original decision to do nothing, without any sign that it had been overturned because the log says it’s the first pass of an administrator. This is considered normal and people don’t complain.”

Question 6: Overall, how much harmful content is on Wikipedia?

Most participants interpreted this question to be asking how much harmful content has made it past the editing process. The consensus is that there is not a lot of content which persists for any length of time, but it does happen rarely and it can sometimes persist for months or years.

Key Quotes

"If one is considering "harmful" from a legal standpoint, I think that there's a fair amount of copyright violating material on Wikipedia, because so few editors deal with it, but I also think that – legally speaking – much of it is justifiable by "fair use", even if it doesn't pass muster by the Wikimedia Foundation's Non-Free Content Criteria (NFCC), which is significantly more stringent than fair use rules."

Question 7: What types of content, if any, do some in the community believe are frequently taken down unnecessarily?

This was a hard question for the participants to answer. There is not a lot of disagreement on Wikipedia about content that gets revision deleted. It is rare to see complaints about the community removing more potentially harmful content than necessary.

The BLP notability guidelines are one of the most common causes of disagreement. Some think the guidelines are applied too strictly, while others argue that the notability guidelines aren't strict enough. One participant noted the controversy around "BLP1E" or "biography of living persons, 1 event". There are some in the community that believe that the moment an individual has been implicated in a crime, they should be talked about on Wikipedia. In this event, the article is sourced entirely from news media, which in turn get info directly from the police. This leads to a situation in which all the information on Wikipedia in this BLP is sourced from the police perspective. BLP1E's are explicitly against policy,[151] but, according to one participant, the enforcement of the policy is inconsistent.

Question 8: What types of content, if any, do some in the community believe should be removed more consistently that currently are not?

According to one participant, some Wikipedians express disappointment when edit summaries that contain harsh criticism of an edit or editor are not deleted. Similarly, there is a lack of consensus on how civility, especially on article talk pages and noticeboards, is policied.

151

https://en.wikipedia.org/wiki/Wikipedia:Biographies_of_living_persons#Subjects_notable_only_for_one_event. Archived at https://perma.cc/YKJ9-ZDUA.

There is content in Australia-related articles that uses offensive terms for indigenous people.

Key Quotes

“Some people have expressed the view that the lack of civility on Wikipedia is destroying the project and driving off editors, that it is the most important problem facing the project,” but others would strongly disagree.

“The last thing we need is an external force, such as WMF, charging in like a bull in a china shop to “fix” what ain’t broke. Our deletion process works well.”

“Rumor has it that the WMF is developing a global civility standard for all of their projects. I think this is a useless and silly thing to attempt to do, considering that – just with the English-speaking community – what’s acceptable in one culture is not acceptable in another. How they’re going to make a standard which crosses all cultures globally, without totally stifling discussion – which is imperative, as Wikipedia runs on consensus, and consensus can only be determined through discussion – I have no idea.”

Question 9: Is there consensus within the community that the current set of guidelines and policies provide an adequate basis for editors to evaluate harmful content? What kind of changes to the guidelines and policies would be helpful?

The participants had mixed opinions about whether there is community consensus about the adequacy of the policies and guidelines.

One editor suspected that if the entire community were asked this question in a survey or poll, the consensus would be that the current set of guidelines do not provide an adequate basis. However, if the community was tasked with creating new or better policies and guidelines, it would fail.

For those editors who said there was consensus among the community, several said that policies are deliberately written and implemented to allow for individual judgement. For some, this is one of the defining features of Wikipedia content moderation, and for others it is a design flaw.

Key Quotes

“Policies are usually deliberately written and implemented to allow individual judgment and consensus to apply foremost, and as long as that's the case I don't think they need changing.”

Question 10: Are the guidelines and policies in regards to removing harmful content interpreted consistently across the community? Why or why not?

The consensus among most of the participants is that the guidelines and policies in regards to harmful content are not interpreted consistently across the community. In unambiguous cases of defamation or clear vandalism, the community does a consistent job. In marginal cases, especially those involving incivility between editors, there are a wide range of opinions and outcomes.

Individual administrators are given a lot of leeway in how they decide to handle issues related to harmful content and conduct. Administrator actions related to removing harmful content are rarely disagreed with or undone without good reason.

Several participants said that the term "harassment" is often misused by editors who break the rules. In other words, editors who have their contributions reverted will accuse the reverting editor of harassing them.[152]

New users are consistently judged more harshly, and veteran users are significantly less likely than new users to be blocked for adding controversial content or for incivility.

According to an oversighter, the more advanced a Wikpedian's permissions are, the more consistent their interpretation and enforcement of the policies and guidelines are with others of the same role. In other words, oversighters are the most consistent group, followed by administrators, with editors being the least consistent. Considering the total number of people within each role, this is unsurprising.

Key Quotes

"No, they are not. First, the community is too large, and second, administrators have discretion about how they interpret policy and behavior, so if things reach the point where an administrator is involved (as an administrator, i.e. for behavioral issues and not content disputes) which administrator acts can make a great deal of difference."

"They are not interpreted consistently. And that is by design...The negative framing is that the guidelines are interpreted inconsistently. But the positive framing is that administrators have autonomy and discretion to handle things the way they think is best."

"Every individual has their own subjective "strike zone," to use a baseball metaphor. What constitutes a ball and a strike is defined and accepted by definition, but there are doubtlessly some who have a slightly wider or tighter "strike zone" than others, just as an aspect of the nature of humanity. This is something that is inevitable."

[152] https://en.wikipedia.org/wiki/Wikipedia:Harassment#Accusing_others_of_harassment. Archived at https://perma.cc/MLP3-PRSL.

“People say new editors and veteran editors should be treated the same…but not necessarily. A person who makes 100,000 appropriate edits and 10 inappropriate edits shouldn’t be treated the same as someone who makes 10 appropriate and 10 inappropriate.”

“The policy of ‘assume good faith’ doesn’t seem to apply towards people on the arbitration committee. Editors often accused arbcom members of being lazy, incompetant, corrupt or biased, and people have no qualms about saying so.”

Question 11: Are there enough editors and administrators to deal with harmful content on Wikipedia?

The opinions among the participants were mixed. Several said that there are enough editors to deal with the amount of harmful content; insufficient editor activity causes problems with routine maintenance rather than with high profile issues. Others said that there are not enough administrators and that administrator burnout is an ongoing issue. The homogeneity of the editor pool is also a concern, more so than the lack of editors overall.

Most participants said that more trusted editors and administrators are always welcome, but that there are more pressing and effective initiatives that the WMF and community can focus on to address harmful content.

The participants noted specific areas within Wikipedia that suffer from a shortage of volunteers, including the OTRS queue and new page patrol.

Key Quotes

“The rewards for removing harmful content based on complaints that show up in the OTRS queue are very small. It is work prone to burnout. You see the worst parts of Wikipedia and worst of people trying to manipulate Wikipedia. It’s terrible work. In an ideal world, paid staff would work on this sort of thing. I wish there were more social standing bestowed upon OTRS volunteers.”

“It’s just a matter of numbers of new page patrollers more than anything. There are currently, I understand, TWO people who are doing more than half of the work there. WMF ultimately treats volunteer editors and administrators as interchangeable, disposable pawns, 100% the “equals” in their glorious site metrics of the drive-bys who make one or two edits a month. They aren’t even tracking the number of “Very Active Editors” (100+ edits per month) now, as opposed to the past, so far as I am aware.”

“In some senses we are always making progress in systemic solutions which result in requiring fewer administrators.”

Question 11a: Would adding additional editors, administrators, or oversighters help to address the amount of harmful content on the platform?

There was not a consensus that adding many new editors would help to address harmful content. A few said that a large influx of new editors would create additional problems.

Key Quotes

“One problem about new editors is that Wikipedia has a pretty steep learning curve, and an editor really shouldn’t be dealing with harmful content issues until they’re familiar enough with the rules, policies, and guidelines to safely navigate them. Too many news editors start contributing and immediately try to be the new sheriff in town, which generally leads to trouble. One thing that could help that is for the WMF to allow for finer gradations of user rights among editors, instead of the 2 or 3 tier system we now have, which is much less stringent than it should be.”

Question 12: Would better mechanisms or automated tools for discovery of harmful content help?

Several participants said that the automated detection tools like Cluebot do a decent job of removing a majority of obvious vandalism and harmful content before it reaches public view. But they are skeptical that automated tools will ever be able to detect and make consistent decisions about the marginal or subtle cases of harm and incivility, considering that the community doesn’t even have a good grasp on what to do about these cases.

Participants expressed an interest in machine learning tools for finding harmful content already existing in articles, as opposed to finding new harmful additions through recent changes.

Key Quotes

“There are existing tools such as the abuse filter, bots, recent changes, and blocking tools, which are under constant development. Long may that continue. Hopefully the watchlist will receive some improvements at some point, as watching things is key.”

“The ability to add an editor’s contribution list to one’s watchlist might be helpful, but I also understand arguments that being able to do so would probably lead to more “following around” an editor and messing with their contributions for spite or to harass them.”

Question 13: Over the previous few questions, we identified various ways in which the content editing process could be improved. What are other obstacles to addressing harmful content and what kinds of approaches and tools would help?

The participants proposed the following issues and/or potential solutions to addressing harmful content on Wikipedia:

- Come to a better community consensus about what constitutes a personal attack.
- Make sure there is a minimum number of editors watching every page.
- The diversity problem of English Wikipedia goes beyond a gender gap. The foundation and community should focus recruitment efforts on addressing diversity with other socioeconomic factors in mind too.
- Create WMF-supported group (or noticeboard/channel, etc) for supporting minority issues and editors. The participant proposing this said it needs to be created and supported by the foundation. Given the lack of minority representation among the editor pool now, a community-organized group would not gain traction.
- Sock puppetry remains a big problem. Elimination of IP editing may be a solution, but editors suggesting this acknowledged how controversial such a change would be.
- Raise the notability guidelines so that there are less BLP pages overall.
- If defamatory material is believed to be a growing problem, make page revisions to BLPs tentative until approved by a trusted editor or administrator.
- Create some constituent policies and guidelines across multiple language projects (other participants explained the difficulties and downsides to this proposal)
- Discretionary sanctions placed on controversial or contentious articles, which are intended to de-escalate disputes, can sometimes get in the way of protecting the article. For example, a one revert rule (1RR) placed on an article makes it so a well-intentioned editor can only protect an article from vandalism one time within the time limit stipulated by the 1RR sanction. If a vandal starts to add harmful content when most US and European editors are asleep, this sanction could prevent an editor from reverting the vandalism if that editor had already used their one allotted revert. If editors accumulated greater editing rights as they continue to work on Wikipedia, discretionary sanctions such as this example could take that into account. Therefore, trusted editors would be able to protect articles in the case suggested above without needing administrator rights.

Key Quotes

“Dynamic IP addresses, or "IP hoppers," are a persistent problem as they often can't be blocked without substantial collateral, if at all. In my opinion we need a way for the Foundation (or community generally) to apply pressure on certain ISPs to get certain abusive users removed from their networks (or get the individuals legally or technically restrained). For example, I would like to know that the Foundation would make a formal complaint about every death threat, or solicitation to murder me, that I regularly receive from ISPs like T-Mobile USA. I have no personal concerns as it happens, but I'm not typical. This is the type of stuff that should be dealt with at that level.”

“The various communities -- American, European, Australian, Indian -- have different beliefs about what is appropriate, what is harmful, and that’s just within English Wikipedia. I don’t know if a uniform code of conduct would ever work.”

“There should be a community process for de-sysopping administrators, something that doesn’t exist now, as well as a required renewal process, with administrators being re-elected every five years. There are pitfalls to that (because administrator actions often make people mad, so there is the strong possibility of revenge voting against renewing administrators), but these can be worked around. I think having to renew their license, so to speak, would help encourage administrators to keep up their skills and take the job seriously.”

Question 14: Of the problems and potential solutions we discussed, which would be most important to prioritize in an attempt to improve how the community addresses harmful content? Why?

The participants gave a wide variety of answers to this question. As a result, there was no clear consensus about one or two initiatives that would be the most important to address. Most of the participants repeated suggestions that they gave for the previous question. Here is a sample of responses that are unique from those above but do not represent a consensus of the most popular answers.

- More avenues and foundational support for community conversation on and off-wiki
- More consistent enforcement of conduct policies
- Better mechanisms to restrain regular vandals
- One participant stressed the need for “clear and unambiguous boundaries about what is considered harmful content within Wikipedia,” with guidance from the WMF

Appendix 2: Current Legal Requirements for Content Moderation in the US, EU, and India.

United States

Due in large part to the expansiveness of the 1st Amendment to the US Constitution,[153] there are only a few legal mechanisms within US law and regulation that address the removal of online content. The primary and most salient of these is Section 230 of the Communications Decency Act, which, by immunizing content hosts and providers from liability for material posted or created by their users, devolves most, if not all, of the control and potential responsibility for removals onto the hosts and providers themselves.

Other partially relevant mechanisms include the notice and take-down provisions of the Digital Millennium Copyright Act (DMCA), court orders for removal predicated on specific areas of US law, such as defamation, and a very small subset of materials that are prima facie to be removed, notably child pornography.

CDA 230

The Communications Decency Act of 1996, a more common name for Title V of the 1996 Telecommunications Act of 1996,[154] is United States federal legislation, much of which was later overturned as unconstitutional.[155] Section 230 (CDA 230), which survived constitutional review and is found at 47 U.S.C. 230, is perhaps the act's most well-known and controversial section,[156] addressing the responsibilities and liabilities for "interactive computer services"—that is, services where users can take action, by posting or uploading material—regarding material on their services. It was included in the law to resolve then-split judicial precedent, with the New York Supreme Court holding in *Stratton Oakmont, Inc*. v. *Prodigy Services Co*.,[157] that online service providers could be held liable for the speech of their users, while the United States District Court for the Southern District of New York in *Cubby, Inc*. v.

---

[153] "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or **abridging the freedom of speech, or of the press;** or the right of the people peaceably to assemble, and to petition the government for a redress of grievances." [emphasis added]

[154] https://www.fcc.gov/general/telecommunications-act-1996. Archived at https://perma.cc/BK42-4728; https://www.govinfo.gov/content/pkg/USCODE-2011-title47/pdf/USCODE-2011-title47-chap5-subchapII-partI-sec230.pdf. Archived at https://perma.cc/7KZF-F8H7; Many pieces of this Act were later struck down as unconstitutional, but CDA 230 has survived. See, e.g. https://www.eff.org/issues/bloggers/legal/liability/230. Archived at https://perma.cc/X3SR-HW4M.

[155] Originally passed in order to "promote competition and reduce regulation in order to secure lower prices and higher quality services for American telecommunications consumers and encourage the rapid deployment of new telecommunications technologies."

[156] 47 U.S.C. § 230.

[157] 1995 WL 323710 (N.Y. Sup. Ct. 1995)

CompuServe Inc.[158] held the opposite, that CompuServe could not be liable because it hadn't moderated.[159]

Most saliently, CDA 230(c) provides the following:

> *(c) Protection for "Good Samaritan" blocking and screening of offensive material*
>
> *(1) Treatment of publisher or speaker*
>
> *No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.*
>
> *(2) Civil liability—No provider or user of an interactive computer service shall be held liable on account of—*
>
> *(A)*
>
> *any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or*
>
> *(B)*
>
> *any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).*

Often misunderstood or misconstrued,[160] this section of the larger act was specifically designed to incentivize online service providers (OSPs) to take steps to moderate or otherwise preview and filter the materials on their service, analogizing to so-called "Good Samaritan" laws indemnifying those offering aid to strangers.[161] In fact, section (b)(4) explicitly states that ["it is the policy of the United States"]

---

[158] 776 F. Supp. 135 (S.D.N.Y. 1991)

[159] For a comprehensive history of CDA 230, see Kosseff, Jeff. *The Twenty-Six Words That Created the Internet*. Cornell University Press, 2019; Kosseff provides a comprehensive list of CDA 230 court opinions and other documents at https://www.jeffkosseff.com/resources. Archived at https://perma.cc/EQ84-M4AH.

[160] See e.g., Mike Masnick. "Nancy Pelosi Joins Ted Cruz And Louis Gohmert In Attacking CDA 230." Techdirt. Accessed May 29, 2019. https://www.techdirt.com/articles/20190411/18521741986/nancy-pelosi-joins-ted-cruz-louis-gohmert-attacking-cda-230.shtml. Archived at https://perma.cc/SA3Y-M2U8; Mike Masnick. "It's One Thing For Trolls And Grandstanding Politicians To Get CDA 230 Wrong, But The Press Shouldn't Help Them." Techdirt. Accessed May 29, 2019. https://www.techdirt.com/articles/20190507/16484342160/one-thing-trolls-grandstanding-politicians-to-get-cda-230-wrong-press-shouldnt-help-them.shtml. Archived at https://perma.cc/6FYE-4EDY; "Inside Facebook's Hellish Two Years—and Mark Zuckerberg's Struggle to Fix It All | WIRED." Accessed May 29, 2019. https://www.wired.com/story/inside-facebook-mark-zuckerberg-2-years-of-hell/.

[161] Specifically those that disfavor imposing a duty on passers-by, in order to make assistance more likely. See, e.g., John T. Pardun,Good Samaritan Laws: A Global Perspective, 20 Loy. L.A. Int'l & Comp. L. Rev. 591 (1998). Available at: http://digitalcommons.lmu.edu/ilr/vol20/iss3/8. Archived at https://perma.cc/D7Q4-H9TP.

"(4) to remove disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material;"

CDA 230 creates these incentives in two distinct ways. First, by making it clear that for any given service, the service itself is not the publisher or "speaker" of any information provided, separating the actions and information of users from any acts of the service and removing the possibility of any sort of contributory liability argument.[162] E.g., YouTube is not the speaker or publisher of a random user comment on a video.

Second, and perhaps even more importantly, CDA 230 makes it possible for service providers to moderate according to their best efforts or ability, as they see fit, without worrying whether their moderation is up to a specific standard. Note the critical language "on account of—any action voluntarily taken in good faith to restrict access to or availability of material." Far from being a blanket immunity that absolves providers from any responsibility to moderate, this instead reveals that the bill's drafters clearly thought what they needed to worry about, and use legislation to pre-empt, was a service provider engaging in some kind of content moderation or removal and subsequently being held liable, either for moderating too much or not enough.

After CDA 230's passage, an interactive online service provider can moderate as much or as little as it sees fit, and as long as the material in question is online by virtue of user action and not by employees of the provider in question, the provider is immunized from liability. Some providers has chosen to engage in little or no moderation of user postings,[163] some more actively, and most somewhere in between, usually with regard to their Terms of Use, often to great controversy.[164]

As written, CDA 230 has "no effect" on federal criminal, intellectual property, or electronic communications privacy law, but state laws inconsistent with CDA 230 cannot be enforced.[165] This latter

[162] Of course, the courts have refined their interpretation of the law since 1996, and it is possible for a service provider to be sufficiently "involved" with a posting to be found to have contributed to its publishing. See, e.g., Fair Housing Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157 (9th Cir. 2008); https://www.eff.org/issues/cda230/cases/fair-housing-council-san-fernando-valley-v-roommatescom. Archived at https://perma.cc/D42J-T399.

[163] E.g., TheDirty.com; See, e.g. Sarah Jones v. Dirty World Entertainment Records LLC, 755 F.3d 398 (6th Cir. 2014) http://www.ca6.uscourts.gov/opinions.pdf/14a0125p-06.pdf. Archived at https://perma.cc/9AMP-LZ8R.

[164] See, e.g., Koebler, Jason, Derek Mead, and Joseph Cox. "Here's How Facebook Is Trying to Moderate Its Two Billion Users." Vice (blog), August 23, 2018; https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works. Archived at https://perma.cc/SA99-UB6Z.
"Onlinecensorship.Org." Accessed June 28, 2018. https://onlinecensorship.org/. Archived at https://perma.cc/DJS3-PY59;
"Twitter to Up Content Moderation as Calls for Regulation Grow | Corporate Counsel." Accessed May 29, 2019; https://www.law.com/corpcounsel/2019/04/16/twitter-to-up-content-moderation-as-calls-for-regulation-grow/. Archived at https://perma.cc/LD36-GUWL.

[165] 47 U.S.C. 230(e)

provision has led to repeated attempts by state-level law enforcement to weaken the law to make state level prosecutions easier.[166]

Most recently, in a deliberate effort to narrow the scope of the the protections that CDA 230 offers to OSPs, the US Congress passed new legislation, the Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA), and the Stop Enabling Sex Traffickers Act (SESTA). Signed into law in April of 2018, this legislation explicitly modified Section 230 to exempt services providers from Section 230 immunity when dealing with civil or criminal crimes related to sex trafficking.[167]

Although CDA 230 has been challenged and misunderstood, and some have called for it to be updated in light of the scale and behaviors of the modern Internet,[168] its tenets and their judicial interpretation are quite clear. Other than with respect to the codified exceptions, any online service provider may moderate, or not, user-generated content as it sees fit, without fear of liability.

Copyright / DMCA Mechanisms for Removal

The Digital Millennium Copyright Act (DMCA), passed in 1998, is the most recent major revision of federal copyright law in the US. Passed in part to address the copyright implications of the then burgeoning Internet, it contains a few key provisions regarding the removal of online material, whether directly, or in removal from search engine results, and creates a "safe harbor" for OSPs who comply with the law's tenets. Critically, the incentives, mechanisms, and penalties within the law, as well as a lack of oversight, make the removal of material after a request extremely likely, which has led to the misuse or abuse of the law's removal requests to achieve the removal of online materials, including harmful speech.

Section 512(c)(3) is the section that addresses requests for removal. Anyone can submit a request, and a request must contain only a few elements.

---

[166] Zimmerman, Matt. "State AGs Ask Congress to Gut Critical CDA 230 Online Speech Protections." Electronic Frontier Foundation, July 24, 2013. https://www.eff.org/deeplinks/2013/07/state-ags-threaten-gut-cda-230-speech-protections. Archived at https://perma.cc/TCG4-4RG3.

[167] The legislation was heavily criticized as being much more about weakening CDA 230 or targeting one particular site, Backpage.com, than actually addressing sex trafficking, and these criticisms have not diminished in the time since the law's passage. See, e.g., "ACLU Letter Opposing SESTA." American Civil Liberties Union. Accessed May 29, 2019. https://www.aclu.org/letter/aclu-letter-opposing-sesta. Archived at https://perma.cc/4X5Y-BHYR; "IMPD Arrests First Suspected Pimp in 7 Months." WRTV, July 3, 2018. https://www.theindychannel.com/longform/running-blind-impd-arrests-first-suspected-pimp-in-7-months. Archived at https://perma.cc/Z73N-X4SR.
Romano, Aja. "A New Law Intended to Curb Sex Trafficking Threatens the Future of the Internet as We Know It." Vox, April 13, 2018. https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom. Archived at https://perma.cc/HT76-6C5W.
"The Scanner: Sex Workers Returned to SF Streets after Backpage.Com Shut down - SFChronicle.Com," October 15, 2018. https://www.sfchronicle.com/crime/article/The-Scanner-Sex-workers-returned-to-SF-streets-13304257.php. Archived at https://perma.cc/A56R-LKM5.

[168] "Jonathan Zittrain » CDA 230 Then and Now: Does Intermediary Immunity Keep the Rest of Us Healthy?" Accessed May 29, 2019. https://blogs.harvard.edu/jzwrites/2018/08/31/cda-230-then-and-now/. Archived at https://perma.cc/Z4GD-QFM8.

(i) A physical or electronic signature of a person authorized to act on behalf of the owner of an exclusive right that is allegedly infringed.

(ii) Identification of the copyrighted work claimed to have been infringed, or, if multiple copyrighted works at a single online site are covered by a single notification, a representative list of such works at that site.

(iii) Identification of the material that is claimed to be infringing or to be the subject of infringing activity and that is to be removed or access to which is to be disabled, and information reasonably sufficient to permit the service provider to locate the material.

(iv) Information reasonably sufficient to permit the service provider to contact the complaining party, such as an address, telephone number, and, if available, an electronic mail address at which the complaining party may be contacted.

(v) A statement that the complaining party has a good faith belief that use of the material in the manner complained of is not authorized by the copyright owner, its agent, or the law.

(vi) A statement that the information in the notification is accurate, and under penalty of perjury, that the complaining party is authorized to act on behalf of the owner of an exclusive right that is allegedly infringed.[169]

A recipient of such a notice will qualify for the safe harbor and be immunized from subsequent liability if it "expeditiously" removes the material that is the subject of the DMCA notice. Critically, this safe harbor will still apply even if turns out that the material was removed erroneously.

DMCA notices are currently sent by the millions annually, and even the largest companies that receive them, such as Google, cannot painstakingly check the provenance of each one. This means that material is often erroneously removed. Sometimes this is because the DMCA notice is a valid one but is overinclusive in what it seeks to remove. However, sometimes it is because the material in question is an inappropriate subject for a removal predicated on copyright, such as commentary or criticism;[170] or while technically appropriate, not a use the law's drafters likely had in mind, such as the removal of non-consensual intimate imagery, or so-called "revenge porn,"[171] which is not currently covered specifically by US federal law, though some companies have their own internal reporting tools for addressing this specifically.[172]

[169] 17 U.S.C 512(c)(3)

[170] See e.g., https://www.theverge.com/2019/4/15/18311091/piracy-tweet-dmca-takedown-request-starz-eff-american-gods. Archived at https://perma.cc/3T6B-34LL.

[171] See, e.g. https://www.dmca.com/faq/How-to-stop-from-being-a-victim-of-revenge-porn. Archived at https://perma.cc/AX93-X6QT; https://www.theatlantic.com/technology/archive/2014/02/our-best-weapon-against-revenge-porn-copyright-law/283564/. Archived at https://perma.cc/5CGM-NRDU.

[172] https://support.google.com/websearch/troubleshooter/3111061#ts=2889054%2C2889099. Archived at https://perma.cc/8EPU-6NT7.

Individual State Laws Addressing Removal

Worthy of notice is the relatively recent proliferation, most likely in the absence of any comprehensive federal legislation, of content removal laws at the state level that are typically aimed at one particular type of content.

The most common of these are laws designed to prevent public law enforcement records being used for extortion,[173] so-called "mugshot" laws, which exist in at least eighteen states, though to questionable effect.[174] More recently California passed a Consumer Privacy Act, which will go into effect in 2020, and which has some parallels with the EU's GDPR, though it has many more exceptions[175]—perhaps too many.[176] New York may soon be following suit with what has been described as an even more stringent[177] privacy law, one that includes the concept of "Information fiduciaries."[178] However, this bill's tenets are also being criticized, this time as unworkable.[179]

Also interesting is California's "right of erasure" law for minors.[180]

---

[173] http://www.abajournal.com/news/article/toptentech. Archived at https://perma.cc/9ZVW-VELN; https://www.legitscript.com/blog/2017/08/mugshots-us-states-that-prohibit-websites-from-charging-a-fee-to-remove-arrest-photos/. Archived at https://perma.cc/7VKK-FCVZ?type=image.

[174] https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2017/12/11/fight-against-mugshot-sites-brings-little-success. Archived at https://perma.cc/7J7Z-2JVU; see also https://en.wikipedia.org/wiki/Mug_shot_publishing_industry#Legislation. Archived at https://perma.cc/Z432-G9FW.

[175] http://ipkitten.blogspot.com/2019/06/california-privacy-law-too-good-to-be.html. Archived at https://perma.cc/PNX2-AE2D.

[176] https://www.techdirt.com/articles/20190120/15122941433/dozens-privacy-experts-tell-california-legislature-that-new-privacy-law-is-badly-undercooked.shtml. Archived at https://perma.cc/6B43-BVX7; https://www.techdirt.com/articles/20180708/00485140195/yes-privacy-is-important-californias-new-privacy-bill-is-unmitigated-disaster-making.shtml. Archived at https://perma.cc/E9P7-T3X2.

[177] https://www.wired.com/story/new-york-privacy-act-bolder/. Archived at https://perma.cc/6UE2-XFLK; https://www.techdirt.com/articles/20190605/07035842338/new-york-states-privacy-law-would-be-among-toughest-us.shtml. Archived at https://perma.cc/VL6A-8VQB.

[178] https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346/. Archived at https://perma.cc/A489-HRF7.

[179] https://www.wired.com/story/new-york-privacy-act-bolder/. Archived at https://perma.cc/6UE2-XFLK; ("'We can't comply with this. We'd have to shut Facebook down in New York,'"

[180] https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201320140SB568. Archived at https://perma.cc/QV88-RL4D. ("The bill would, on and after January 1, 2015, require the operator of an Internet Web site, online service, online application, or mobile application to permit a minor, who is a registered user of the operator's Internet Web site, online service, online application, or mobile application, to remove, or to request and obtain removal of, content or information posted on the operator's Internet Web site, service, or application by the minor")

European Union

The Copyright Directive and Current Revisions

Somewhat analogously to the United States' CDA 230, the EU's Directive 2000/31/EC[181] sometimes known as the e-commerce directive, establishes a type of safe haven regime for hosting providers:

- Article 14 establishes that hosting providers are not responsible for the content they host as long as (1) the acts in question are neutral intermediary acts of a mere technical, automatic and passive capacity; (2) they are not informed of its illegal character, and (3) they act promptly to remove or disable access to the material when informed of it.
- Article 15 precludes member states from imposing general obligations to monitor hosted content for potential illegal activities.

However, this Directive is currently in the process of being modified. Article 13 of the Directive on Copyright in the Digital Single Market,[182] which in April of 2019 was passed on a first reading, but as of this writing has not yet ratified by the Council of Europe's ministers,[183] would make providers liable if they fail to take "effective and proportionate measures" to prevent users from uploading certain copyright violations and do not respond immediately to takedown requests.

GDPR

As of a European Union Court of Justice (EUCJ) ruling on the *Google Spain* v. *AEPD and Mario Costeja González* case in May of 2014, the major relevant tenets of which were subsequently codified into law in 2018 as the General Data Protection Regulation (GDPR), online content in EU member countries[184] is subject to possible removal under what is typically known as the "right to be forgotten" but is more accurately called the "right to erasure" under the GDPR.[185]

Individuals or entities making such a request to a data controller or processor need only argue that the information in question meets one of six criteria.[186] Upon receipt of such a request, the controller or

[181] https://en.wikipedia.org/wiki/Directive_2000/31/EC. Archived at https://perma.cc/8CSB-YCXL.

[182] https://en.wikipedia.org/wiki/Directive_on_Copyright_in_the_Digital_Single_Market. Archived at https://perma.cc/2AU2-JR7J.

[183] See, e.g., https://www.cnbc.com/2019/03/26/eu-parliament-passes-copyright-ruling-that-will-hit-google-facebook.html. Archived at https://perma.cc/36ZT-KNTR. If the ministers ratify the legislation, member states will from that point forward have two years to implement the law.

[184] And arguably worldwide, see, e.g. http://curia.europa.eu/juris/document/document.jsf?text=&docid=214686&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=7255506. Archived at https://perma.cc/V2HS-C63B;

[185] https://gdpr-info.eu/art-17-gdpr/. Archived at https://perma.cc/4MJ6-5B5D.

[186] "the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed; the data subject withdraws consent on which the processing is based according to point (a) of Article 6(1), or point (a) of Article 9(2), and where there is no other legal ground for the processing; the data subject objects to the processing pursuant to Article 21(1) and there are no overriding legitimate grounds for the

processor may choose to comply or not, and denial of a request may be appealed to the member state's data protection authority (DPA).

Code of Conduct: Hate Speech on Online Platforms

Another possible removal mechanism is specifically with respect to illegal hate speech[187] found on online platforms. In 2016, the European Commission, in conjunction with Facebook, Microsoft, Twitter, and YouTube,[188] published a (voluntary) Code of Conduct which has, as of the spring of 2019, been evaluated four times,[189] with five more companies signing on to adhere to the code. Monitoring as to successful adherence to the code is carried out by "a network of civil society organisations located in different EU countries." Using a commonly agreed methodology, these organisations test how the IT companies apply the Code of Conduct in practice."[190]

The most recent evaluation found that

> "IT companies are now assessing 89% of flagged content within 24 hours and 72% of the content deemed to be illegal hate speech is removed, compared to 40% and 28% respectively when the Code[191] was first launched in 2016."[192]

For a much more comprehensive recent review of EU hate speech law, in six of its member countries, see "Responding to 'hate speech': A comparative overview of six EU countries" from the Article19 organization.[193]

Worthy of a brief note is the late June 2019 announcement from Facebook that it would begin to share the IP addresses and other identification data of individuals suspected of hate speech to French judges

processing, or the data subject objects to the processing pursuant to Article 21(2); the personal data have been unlawfully processed; the personal data have to be erased for compliance with a legal obligation in Union or Member State law to which the controller is subject; the personal data have been collected in relation to the offer of information society services referred to in Article 8(1)."

[187] Defined in EU law under the Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law as the "public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178. Archived at https://perma.cc/ZVP2-ASK3.

[188] http://europa.eu/rapid/press-release_MEMO-19-806_en.htm. Archived at https://perma.cc/76PA-NQWD.

[189] Id.

[190] Id.

[191] More information from the EU on the history, mechanisms, and successes of the Code is available at https://ec.europa.eu/info/files/factsheet-how-code-conduct-helped-countering-illegal-hate-speech_en. Archived at https://perma.cc/365W-X25Q; http://europa.eu/rapid/press-release_IP-19-805_en.htm. Archived at https://perma.cc/3G3J-L78T.

[192] https://ec.europa.eu/commission/news/countering-illegal-hate-speech-online-2019-feb-04_en. Archived at https://perma.cc/B9PT-4WMZ.

[193] https://www.article19.org/wp-content/uploads/2018/03/ECA-hate-speech-compilation-report_March-2018.pdf. Archived at https://perma.cc/XWJ2-Q35X.

who formally demanded it.[194] Facebook had previous refrained from so doing "because it was not compelled to do so under U.S.-French legal conventions and because it was worried countries without an independent judiciary could abuse it."[195]

Illegal Content on Online Platforms

The EU does not yet have a fully instantiated policy regarding the broad category of "illegal content located online," but on March 1, 2018 published a set of recommendations on this topic designed to "effectively tackle illegal content online."[196] These recommendations build on earlier efforts from 2017, and translate those earlier political efforts into (non-binding) legal form.[197] The recommendations are still at a quite broad level, with no concrete specifics, but suggest that online platforms should be more responsible in general, in order to "swiftly and proactively detect, remove and prevent the reappearance" of illegal content.[198] The five high-level recommendations are:

- Clearer 'notice and action' procedures
- More efficient tools and proactive technologies
- Stronger safeguards to ensure fundamental rights
- Special attention to small companies
- Closer cooperation with authorities

Notably, the EU's 2018 assertion that "online intermediaries can put in place proactive measures without fearing to lose the liability exemption under the e-Commerce Directive" should be considered in light of a recent June 2019 opinion from EU Advocate General Szpunar, issued regarding Case C-18/18, *Eva Glawischnig-Piesczek v. Facebook Ireland Limited*, in Section 2(a) of which he discusses the possibility that imposing any kind of general monitoring obligation might mean that an online platform could lose its Article 14 safe harbor immunity;[199] as well as in light of the October 3, 2019 final ruling,

---

[194] https://www.reuters.com/article/us-france-tech-exclusive/exclusive-facebook-to-give-data-on-hate-speech-suspects-to-french-courts-minister-idUSKCN1TQ1TJ. Archived at https://perma.cc/5EGJ-UY5D.

[195] Id.

[196] https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms. Archived at https://perma.cc/4Q6E-4PL9.

[197] Id.

[198] https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms. Archived at https://perma.cc/4Q6E-4PL9.

[199] "If, contrary to that provision, a Member State were able, in the context of an injunction, to impose a general monitoring obligation on a host provider, it cannot be precluded that the latter might well lose the status of intermediary service provider and the immunity that goes with it" http://curia.europa.eu/juris/document/document.jsf?text=&docid=214686&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=7255506. Archived at https://perma.cc/LJ6Z-YGS8; for a longer and perhaps U.S.-centric discussion of the AG Szpunar's opinion and its implications for content removal, see https://www.techdirt.com/articles/20190604/13305042334/european-court-justice-suggests-maybe-entire-internet-should-be-censored-filtered.shtml. Archived at https://perma.cc/958Q-YSBX.

Eva Glawischnig-Piesczek v. Facebook Ireland Limited,[200] where the EUCJ ruled against Facebook, holding that after receiving a court order to remove allegedly defamatory material regarding an Austrian politician, Facebook was not only required to do so but had to do so globally.

"Terrorist Content" Online

In April of 2019, the EU Parliament voted in favor of[201] a measure regarding online platforms' "responsibilities" for addressing and removing so-called "terrorist content." The definition of what qualifies as such content is complex, and there are also several important exceptions.[202] "The newly elected European Parliament will be in charge of negotiating with the Council of Ministers on the final form of the text."[203] Companies have no obligation to monitor generally (see again AG Szpunar's opinion) but must remove content within one hour upon receiving notification or face fines of up to 4% of revenue. The current version of the measure does not contain some of the original's more controversial and difficult to implement features,[204] but even in this less restrictive form has been the subject of substantial criticism.[205]

Country-specific Removal Laws

Within the EU, member states can and have passed their own internal laws and regulations concerning the removal of online content. In addition to the Austrian defamation law that is at the heart of the

200 http://curia.europa.eu/juris/document/document.jsf?text=&docid=218621&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=1965965. Archived at https://perma.cc/VZ8E-DFKR.

201 http://www.europarl.europa.eu/news/en/press-room/20190410IPR37571/terrorist-content-online-should-be-removed-within-one-hour-says-ep. Archived at https://perma.cc/GQ6R-E5CQ.

202 "The legislation targets any material—text, images, sound recordings or videos—that "incites or solicits the commission or contribution to the commission of terrorist offences, provides instructions for the commission of such offences or solicits the participation in activities of a terrorist group," as well as content providing guidance on how to make and use explosives, firearms and other weapons for terrorist purposes.
Content disseminated for educational, journalistic, or research purposes should be protected, according to MEPs. They also make clear that the expression of polemic or controversial views on sensitive political questions should not be considered terrorist content." http://www.europarl.europa.eu/news/en/press-room/20190410IPR37571/terrorist-content-online-should-be-removed-within-one-hour-says-ep. Archived at https://perma.cc/GQ6R-E5CQ.

203 http://www.europarl.europa.eu/news/en/press-room/20190410IPR37571/terrorist-content-online-should-be-removed-within-one-hour-says-ep. Archived at https://perma.cc/GQ6R-E5CQ; an earlier version can be found at https://ec.europa.eu/commission/sites/beta-political/files/soteu2018-preventing-terrorist-content-online-regulation-640_en.pdf. Archived at https://perma.cc/MFJ4-QZWP.

204 https://edri.org/eu-parliament-deletes-worst-threats-to-freedom-of-expression-terrorist-content-regulation. Archived at https://perma.cc/ZYN8-BW8X.

205 https://www.theverge.com/2019/3/21/18274201/european-terrorist-content-regulation-extremist-terreg-upload-filter-one-hour-takedown-eu. Archived at https://perma.cc/QQD8-PGCH; https://edri.org/eu-parliament-deletes-worst-threats-to-freedom-of-expression-terrorist-content-regulation/. Archived at https://perma.cc/ZYN8-BW8X.

previously-mentioned *Eva Glawischnig-Piesczek v. Facebook Ireland Limited* case, there are others, including Germany's "NetzDG" law,[206] which took effect on January 1, 2018.

This law requires social networks such as Facebook to remove material that violates aspects of German law, most notably including broad prohibitions on "defamation of religion," "hate speech," and "insult." Controversially,[207] the removal obligation stems from complaints by users, and the material at issue must be removed within 24 hours or seven days depending on its severity, with the company facing steep fines for a failure to do so.[208] Related to the fact that it is users, not law enforcement or a court, that can submit a complaint is that there is no way to challenge a complaint or removal,[209] and that the social media companies themselves become the arbiters of what is and is not qualifying content.[210] Some interesting statistics on NetzDG removals have already been collected,[211] and at least three other countries, including some not in the EU, are pursuing the passage of similar legislation using NetzDG as a model.[212]

India

"Hate" or Otherwise Objectionable Speech

Indian law does not use the phrase "hate speech." Different forms of what may arguably be called hate speech are covered in different ways by various Indian statutes,[213] and research on this topic typically refers to the speech in question as "hate speech."

---

[206] https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=80C18C3BD 91D8DE28FB6AC2317FA4A28.1_cid289?__blob=publicationFile&v=2. Archived at https://perma.cc/3GNU-Z2DH.

[207] https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law. Archived at https://perma.cc/B5R3-993G.

[208] https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=80C18C3BD91D8DE28FB6AC2317FA4A28.1_cid289?__blob=publicationFile&v=2. Archived at https://perma.cc/3GNU-Z2DH; https://www.justsecurity.org/62857/eu-terrorist-content-proposal-sets-dire-free-speech-online/. Archived at

[209] https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law. Archived at https://perma.cc/B5R3-993G; https://www.justsecurity.org/62857/eu-terrorist-content-proposal-sets-dire-free-speech-online/. Archived at https://perma.cc/E2HQ-VTEB; https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=80C18C3BD91D8DE28FB6AC2317FA4A28.1_cid289?__blob=publicationFile&v=2. Archived at https://perma.cc/3GNU-Z2DH.

[210] https://www.theatlantic.com/international/archive/2018/05/germany-facebook-afd/560435/. Archived at https://perma.cc/URA3-KGHB.

[211] https://transparencyreport.google.com/netzdg/youtube?hl=en. Archived at https://perma.cc/KG7T-EGB4

[212] https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law. Archived at https://perma.cc/B5R3-993G.

[213] Arun, Chinmayi, and Nayak, Nakul. Preliminary Findings on Online Hate Speech and the Law in India (December 8, 2016). Berkman Klein Center Research Publication No. 2016-19. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882238. Archived at https://perma.cc/VLM3-Y5L5.

Section 95 of the Indian Code of Criminal Procedure, 1973 (CrPC), in which larger document the majority of the regulation of speech is found, defines objectionable speech as "matter which promotes or is intended to promote feelings of enmity or hatred between different classes of citizens."[214] "Section 196 of the CrPC acts as a procedural safeguard against frivolous prosecution for 'hate speech' offences, such as those under Sections 153A, 295A, 505 and 153B of the IPC."[215] These definitions and limitations largely carry over to online incidences of hate speech, although there is a range of applicable laws and enforcement mechanisms are more online-specific.

Broadly, there are three ways in which online "hate speech" can be addressed within India law and regulation.[216] The speech content in question can be blocked, filtered or removed entirely; criminal action can be taken against the author or intermediary; and finally, Internet access itself (to the content in question) can be blocked.[217] This governmental power has been reviewed and found constitutional as recently as 2015.[218]

Blocking, Filtering and Removal

Section 69A of the IT act covers blocking of objectionable material. "The procedure to be followed for blocking of access is contained in the Information Technology (Procedure and Safeguards for Blocking for Access of Information by Public) Rules, 2009 (Blocking Rules)."[219] The government can issue a directive that access to material be blocked, on one or more of six possible grounds (most typically "public order") and failure to comply can result in a fine and an up to seven-year prison sentence.[220]

Removal of Online Content and the Safe Harbor Law

India has what is typically referred to as a "safe harbor " law. More specifically, this is Section 79 of the Information Technology Act ("IT Act"),[221] (amended most recently in 2008) and its accompanying guidelines for intermediaries.[222] In a nutshell, this law is somewhat similar to the United States CDA 230

[214] https://indiankanoon.org/doc/875455/. Archived at https://perma.cc/8AYX-DPGX.
[215] Center For Communication Governance, "Hate Speech Laws in India" p.66; https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view.
[216] Center For Communication Governance, "Hate Speech Laws in India" p.133; https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view; See also Report No. 267 - Law Commission of India. March 2017. http://lawcommissionofindia.nic.in/reports/Report267.pdf. Archived at https://perma.cc/4SHW-B2RS. for both a more extensive examination of current jurisprudence on hate speech in India and its effects on freedom of expression; as well as international comparisons.
[217] Id.
[218] For more details on the mechanisms and contours of blocking under Section 69A, see pp. 136-138, 7.3.x of Center For Communication Governance, "Hate Speech Laws in India" .
[219] Center For Communication Governance, "Hate Speech Laws in India" p.136; https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view.
[220] Id.
[221] https://cis-india.org/internet-governance/resources/section-79-information-technology-act. Archived at https://perma.cc/A9PJ-PFVH.
[222] https://www.meity.gov.in/writereaddata/files/Draft_Intermediary_Amendment_24122018.pdf. Archived at https://perma.cc/HQ2U-2AFT.

(and other roughly analogous legislation) in that under certain circumstances, it immunizes online intermediaries from liability for the content created by third parties even when the content is against other Indian laws.

Broadly, the requisite factors are that the intermediary not be actively and deliberately involved with the third party's posting and exercises due diligence;[223] and that once notified of the illegality of a posting, the intermediary act expeditiously to remove it.[224]

> "The Supreme Court has held that intermediaries are only required to take down content 'upon receiving actual knowledge from a court order, or on being notified by the appropriate government or its agency and not on the basis of user complaints."[emphasis added][225]

Importantly, this means that user notification alone is not enough to impute knowledge (that would lead to liability) to an online platform.

Criminalization

Although Section 66A of the IT Act originally contained other and more far-reaching provisions that allowed for the criminalization of online speech; and Section 79 was more powerful in its application, a 2015 ruling from the Supreme Court of India in *Shreya Singhal v. Union of India*[226] narrowed the scope of the act by, among other things, deciding that Section 66A was unconstitutional, given the degree to which it interfered with the freedom of speech guaranteed by the Indian Constitution; and that Section 79 and its rules would only apply if the intermediary in question received a court order. The ruling and its effect on the law and its application has been described as a key event for the future of Indian free speech.[227]

As a result of this ruling, intermediaries hosting online hate speech are not currently subject to criminal proceedings, although they may have been in the past. There are currently a variety of proposals to

[223] Section 79(2)(b) of the IT Act

[224] Compare, for example, the DMCA's "remove expeditiously" language and notification provisions, and the case law surrounding CDA 230's definitions of a third party, and an intermediaries "involvement" or "inducement. See e.g. https://ilt.eff.org/Defamation__CDA_Cases.html#Internet_Service_Provider_v._Internet_Content_Provide. Archived at https://perma.cc/6GKR-GC3D.

[225] Center For Communication Governance, "Hate Speech Laws in India" p. 139 (internal citations omitted); https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view

[226] For expert discussion of this case, see, e.g., https://www.thehindu.com/opinion/op-ed/shreya-singhal-case-of-the-online-intermediary/article7074431.ece. Archived at https://perma.cc/2PEM-DQCV.

[227] See, e.g., https://indconlawphil.wordpress.com/2015/03/26/the-striking-down-of-section-66a-how-indian-free-speech-jurisprudence-found-its-soul-again/. Archived at https://perma.cc/9JRQ-YQ4T. ("So perhaps, at long last, the time has come to rethink fifty-year old judgments upholding blasphemy and sedition laws, rethink criminal defamation, throw off the oppressive fetters of civil defamation and contempt of court, and attack the censorship guidelines of both cinema and cable TV."

further alter or update the IT Act to make it more in harmony with Sections 153A and 153B of the Indian Penal Code.[228]

Also worth a brief mention are some recently proposed changes to Section 79,[229] which are intended to address the spread of "fake news" and other putatively undesirable content on social networks. The proposed changes, in addition to requiring applications such as WhatsApp to decrypt encrypted data upon government request, would require online intermediaries to implement automatic "filtering" techniques to prevent objectionable materials from ever being posted to begin with. The Indian government has yet to comprehensively define what material would be covered by the new law, and although some companies are preemptively removing or disallowing certain materials in anticipation, the proposed measures are being sharply criticized as an assault on freedom of speech. Some Indian legal experts doubt that they will ever pass.

N.B. India does also have a copyright-based safe harbor separate and distinct from the "safe harbor" law described above. This version of a safe harbor is found in the Indian Copyright Act, originally written in 1957 and most recently amended in 2012, with a 2013 set of accompanying rules.[230]

Internet Shutdowns (Total Lack of Access)

It is also possible (and increasingly frequent) under some circumstances for the Indian government to issue an order that will "shutdown all access to information in response to perceived threats to law and order."[231] Such shutdowns are typically ordered under the authority of Section 144 of the CrPC and the accompanying 'Temporary Suspension of Telecom Services (Public Emergency or Safety) Rules, 2017'; and the Telegraph Act of 1957.[232]

District Magistrates can issue such orders if they are based on material facts and "absolute and definite" in their terms. Further, such orders must be for only "immediate prevention" and can only last for the duration of the stated emergency.[233] According to rules issued in August of 2017, "only the 'Secretary to the Government of India in the Ministry of Home Affairs in the case of Government of India or the Secretary to the State Government in-charge of the Home Department in the case of a State Government' can issue directions for shutdowns."[234] Such shutdowns, thought quite frequent, have been challenged several times, and have always been found to be constitutional based on the required

[228] Center For Communication Governance, "Hate Speech Laws in India" p. 135. https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view
[229] https://www.meity.gov.in/writereaddata/files/Draft_Intermediary_Amendment_24122018.pdf. Archived at https://perma.cc/HQ2U-2AFT.
[230] http://copyright.gov.in/Documents/Copyright_Rules_2013_and_Forms.pdf. Archived at https://perma.cc/N2RT-LZAC.
[231] Center For Communication Governance, "Hate Speech Laws in India" p. 141. https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view
[232] Id.
[233] Id.
[234] Id.

urgency as well as other limitations. However, "the process of issuing blocking orders is opaque, and the reasoning offered in orders is not subject to public scrutiny."[235]

In the late summer and fall of 2019, the Indian government initiated just such a region-wide shutdown in Kashmir (as did Pakistan).[236] The Kashmir region accounts for 70% of the Indian government's Internet shutdowns, and this latest was the 133rd such shutdown of 2019.[237] The lengthy shutdown has led to criticism and protest from journalists,[238] doctors,[239] human rights groups, [240] and the UN.[241] It has also led to an increase in civic ignorance[242] and caused substantial estimated losses to the local economy.[243]

[235] Center For Communication Governance, "Hate Speech Laws in India" p. 142. https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view

[236] https://www.nytimes.com/2019/08/14/technology/india-kashmir-internet.html; Archived at https://perma.cc/5JPH-VPUU. https://www.washingtonpost.com/world/internet-mobile-blackout-shuts-down-communication-with-kashmir/2019/08/06/346d5150-b7c4-11e9-8e83-4e6687e99814_story.html; Archived at https://perma.cc/EYW4-6FEP. https://netblocks.org/reports/pakistan-shuts-down-internet-in-kashmir-restricts-access-in-punjab-and-beyond-3Anw7dB2. Archived at https://perma.cc/2QBM-F7GG.

[237] https://internetshutdowns.in/. Archived at https://perma.cc/CEP7-JJF9.

[238] https://www.indiatoday.in/india/story/kashmir-lockdown-journalists-protest-clampdown-demand-mobile-phones-internet-1605920-2019-10-03. Archived at https://perma.cc/HDK9-PTL5.

[239] https://www.youtube.com/watch?reload=9&v=jK3lgRAqgPk.

[240] https://www.hrw.org/news/2019/08/28/india-restore-kashmirs-internet-phones#. Archived at https://perma.cc/RJ4A-S58D.

[241] https://www.theguardian.com/world/2019/aug/08/kashmir-communications-blackout-is-draconian-says-un-envoy. Archived at https://perma.cc/B2D8-YKB7.

[242] https://www.businessinsider.com/kashmir-blackout-locals-dont-know-india-revoked-autonomy-report-2019-8. Archived at https://perma.cc/9MUT-5BZJ.

[243] https://www.hrw.org/news/2019/08/28/india-restore-kashmirs-internet-phones#. Archived at https://perma.cc/RJ4A-S58D.

Appendix 3: Quantitative Methods

Most of our work involved counting, for which the Quarry tool was exceptionally useful.[244] Our queries were run between April and July 2019.

To extract text added to a page as part of a revision, we used Python's difflib library to find all added character spans between a revision and its parent. Each revision could then contribute multiple character spans plus the username and revision summary. Each of these pieces of content were tested for harmful content, and if any were classified as harmful, the revision counted as harmful.

The bulk of our analysis using the Perspective API took place between July 23 and July 29, 2019. To use the Perspective API to locate harmful content, we applied the following rules to text. First, the piece of text must be longer than 5 characters. This helped eliminate small texts that scored highly but were unlikely to be harmful on their own, e.g. "kill". The text must then be classified by the Perspective API as "severely toxic" with 90% confidence or greater using version 2 of the English "SEVERE_TOXICITY" model, a "threat" with 93% confidence or greater using version 2 of the experimental English "THREAT" model, or an "identity-based attack" with 90% confidence or greater using version 2 of the experimental English "IDENTITY_ATTACK" model. These thresholds were determined by manually reviewing how well they separated truly harmful content from benign content that appeared to confuse the Perspective classifiers. The threat and identity-based attack classifiers were picked from the 16 potential classifiers because these two types of content are consistently discussed by legislative bodies around the world. In an effort to catch other content which may not be a threat or identity-based attack but might still be considered harmful in a legal sense, we added the "severely toxic" classifier to our process.

Across the four primary namespaces, Perspective flagged 317 of 400,000 revisions as including harmful content. We manually reviewed these 317 revisions and removed obvious false positives. "Obvious" here means anything that we deemed very unlikely to be seen as harmful. For example, one revision added the words "and kill them" to a sentence about a scheme of the Penguin in the 1992 movie Batman Returns and was subsequently flagged as harmful. After removing these false positives, we were left with 238 revisions. This means the process outlined above had about a 25% false positive rate.

In an effort to estimate the false negative rate, we randomly sampled from revisions that had been tagged as possible vandalism by the tag filtering system in place on English Wikipedia. We deemed it unrealistic to review a random sample of all negatively classified revisions in an attempt to locate false negatives given the very low prevalence of detectable harmful content. Sampling from a population of revisions likely to contain a higher prevalence of harmful content does not help us estimate an overall false negative rate, but it does provide a quick (but soft) upper bound. After reviewing 100 randomly sampled revisions that had been tagged as possible vandalism but had not been classified as harmful, we identified two that we would expect the process outlined above to detect. Additionally, we identified 20 more that would fall into our taxonomy of harmful content. Almost all of these would be considered defamation.

[244] https://quarry.wmflabs.org/

To collect data on the lifetime of harmful content, for each piece of harmful content within each revision, we stepped forward in time through subsequent revisions and found the first revision which no longer contained the text that was classified as harmful. To determine the lifetime, we took the difference between the creation dates of the original revision and this revision. The editors who performed these reverts informed our analysis of who removes harmful content.

To determine the amount of harmful content in the objectionable content, controversial topics, and tagged vandalism revision spaces, we collected all the revision IDs in each of these spaces and randomly sampled from those sets. The resulting revisions were run through the same process described above.

Appendix 4: Glossary

Administrator: Wikipedia editor who have been granted advanced technical permissions so that they can perform the following actions: Revision deletion, block and unblock users and IP addresses, delete and protect pages from editing, and grant user permissions.[245] [246]

CheckUser: Wikipedia editor with advanced permissions that allow them to determine the IP addresses of Wikipedia accounts. Checkusers use this tool to investigate and prevent cases of disruptive editing or sock puppetry. [247]

Edit: We use this term to refer to any changes (additions or removals) to any content on Wikipedia. In this report, the term edit is used interchangeably with the term revision. We were unable to find a definition on Wikipedia.

Edit Filter: An edit filter automatically compares every edit made to Wikipedia against a defined set of conditions. If an edit matches the conditions of a filter, that filter will respond by logging the edit. It may also tag the edit summary, warn the editor, revoke his/her autoconfirmed status, and/or disallow the edit entirely.[248]

Editor: Anyone who writes or modifies Wikipedia or otherwise contributes to the platform.[249]

Edit Summary: A brief explanation of an edit to a Wikipedia page.[250]

Guideline: Best practice for following policies in specific contexts. See also "Policy"[251]

Namespace: A Wikipedia namespace is a set of Wikipedia pages whose names begin with a particular reserved word recognized by the MediaWiki software (followed by a colon). For example, in the user namespace all titles begin with the prefix "User:"[252]

Noticeboard: A page that acts as a forum for a group of users, who use it to coordinate their editing. Most notice boards are by geographic location, like the UK Wikipedians' notice board; a notable exception is the Administrators' noticeboard.[253]

[245] https://en.wikipedia.org/wiki/Wikipedia:Administrators. Archived at https://perma.cc/5LRC-MYQS.
[246] https://en.wikipedia.org/wiki/Wikipedia:Requests_for_permissions. Archived at https://perma.cc/9LTV-L58G.
[247] https://en.wikipedia.org/wiki/Wikipedia:CheckUser. Archived at https://perma.cc/8MNE-LZKE.
[248] https://en.wikipedia.org/wiki/Wikipedia:Edit_filter. Archived at https://perma.cc/ZWN2-3UA5.
[249] https://en.wikipedia.org/wiki/Wikipedia:Glossary#Editor. Archived at https://perma.cc/TS7A-EYGN.
[250] https://en.wikipedia.org/wiki/Help:Edit_summary. Archived at https://perma.cc/QVZ9-VGKE.
[251] https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines. Archived at https://perma.cc/W9U6-7EAJ.
[252] https://en.wikipedia.org/wiki/Wikipedia:Namespace. Archived at https://perma.cc/9LRM-9V2T
[253] https://en.wikipedia.org/wiki/Wikipedia:Glossary#Noticeboard. Archived at https://perma.cc/TS7A-EYGN.

ORES: A web service and API that provides machine learning as a service. The system is designed to help automate critical wiki-work—for example, vandalism detection and removal. Currently, the two general types of scores that ORES generates are in the context of "edit quality" and "article quality."[254]

OTRS: Open Ticket Request System. Refers to the people and software that surround the handling of email sent to the Wikimedia Foundation.[255]

Oversight: See "Suppression"

Oversighter: Wikipedia editor with advanced permissions that allow them to perform the following actions: suppress or unsuppress revisions or log entries, suppress an account user name in conjunction with an account block, review all suppressed revisions or log entries.[256]

Policy: Standards that all users should normally follow. Divided into "Content" and "Conduct" policies.[257]

Revert: An edit that reverses edits made by someone else, thus restoring the prior version.[258]

Revision: See "Edit"

Revision Deletion: RevisionDelete (also known as RevDel or RevDelete) is a feature that allows administrators to remove individual entries in a page history or log from public view.[259]

Sock Puppet: Another user account created secretly by an existing Wikipedian, generally to manufacture the illusion of support in a vote or argument.[260]

Suppression: Suppression on Wikipedia (also known as oversight for historical reasons) is a form of enhanced deletion that, unlike normal deletion, expunges information from any form of usual access, even by administrators.[261]

Tag: A mediawiki tag, brief message applied next to certain revisions by the software.[262]

Template: A way of automatically including the contents of one page within another page, used for boilerplate text, navigational aids, etc.[263]

---

[254] https://www.mediawiki.org/wiki/ORES. Archived at https://perma.cc/R8B7-LWT7.
[255] https://en.wikipedia.org/wiki/Wikipedia:Glossary#Open_Ticket_Request_System. Archived at https://perma.cc/TS7A-EYGN.
[256] https://en.wikipedia.org/wiki/Wikipedia:Oversight. Archived at https://perma.cc/TJ6M-TBLG.
[257] https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines. Archived at https://perma.cc/W9U6-7EAJ.
[258] https://en.wikipedia.org/wiki/Wikipedia:Glossary#Revert. Archived at https://perma.cc/TS7A-EYGN.
[259] https://en.wikipedia.org/wiki/Wikipedia:Revision_deletion. Archived at https://perma.cc/6C5F-TSSC.
[260] https://en.wikipedia.org/wiki/Wikipedia:Glossary#Sock_puppet. Archived at https://perma.cc/TS7A-EYGN.
[261] https://en.wikipedia.org/wiki/Wikipedia:Oversight. Archived at https://perma.cc/TJ6M-TBLG.
[262] https://en.wikipedia.org/wiki/Wikipedia:Glossary#Tag. Archived at https://perma.cc/TS7A-EYGN.
[263] https://en.wikipedia.org/wiki/Wikipedia:Glossary#Template. Archived at https://perma.cc/TS7A-EYGN.

Vandalism: Editing (or other behavior) deliberately intended to obstruct or defeat the project's purpose, which is to create a free encyclopedia, in a variety of languages, presenting the sum of all human knowledge.[264]

Watchlist: A watchlist is a page which allows any logged-in user to keep a list of "watched" pages and to generate a list of recent changes made to those pages.[265]

[264] https://en.wikipedia.org/wiki/Wikipedia:Vandalism. Archived at https://perma.cc/8N4Z-WKD3.
[265] https://en.wikipedia.org/wiki/Help:Watchlist. Archived at https://perma.cc/J6MK-2EJP.