Commentary

Assuming Good Faith Online

Eric Goldman

1 Introduction

One of Wikipedia's fundamental principles is to "assume good faith" (Wikipedia 2024a). The Wikipedia project page explains: "It is the assumption that editors' edits and comments are made in good faith—that is, the assumption that people are not deliberately trying to hurt Wikipedia, even when their actions are harmful." (Wikipedia 2024a; Reagle 2010; Ayers et al. 2008).

In theory, this principle should not be remarkable. Most people act in good faith most of the time (Bregman 2020)—and any well-functioning society depends on good-faith interactions as the norm. Nevertheless, Wikipedia's assume-good-faith principle feels remarkable because it defies the widespread and obvious evidence that Internet users routinely engage in bad-faith activity on Wikipedia. A service as large and visible as Wikipedia inevitably attracts users who advance their self-interest to the disadvantage of Wikipedia's contributors and readers (Goldman 2010).

Wikipedia's assume-good-faith principle "does not require that editors continue to assume good faith in the presence of obvious evidence to the contrary" (Wikipedia 2024a). In fact, the Wikipedia community will sometimes moderate content without having obvious evidence in hand. For example, Wikipedia relies heavily on bots (Nasaw 2012; Wikipedia 2023). "Bots are playing an increasingly important role in the creation of knowledge in Wikipedia" (Zheng et al. 2019), even if the bots cannot "assume" good faith.[1] Furthermore, the Wikipedia community has developed xenophobic tendencies (Goldman 2010) that are skeptical of newcomers' activities, even before the newcomers demonstrate any bad faith.[2] Perhaps weary after two decades of fighting bad-faith activity, the Wikipedia community struggles to retain its foundational assumption of good faith.

Wikipedia's tension between assuming users' good faith while combating users' bad-faith contributions is not unique. Every Internet service[3] enabling user-generated content faces a similar challenge to balance good-faith and bad-faith activity. Without a proper balance, a service foregoes one of the Internet's signature features—users' ability to engage with and learn from each other in pro-social and self-actualizing ways—and instead pushes toward one of two suboptimal outcomes. Either the service devolves into a cesspool of bad-faith activity, or it becomes a restrictive and locked-down venue with limited expressive options for all users, including the well-intentioned ones.

1. For example: "Although Wikipedia bots are intended to support the encyclopedia, they often undo each other's edits and these sterile 'fights' may sometimes continue for years. Unlike humans on Wikipedia, bots' interactions tend to occur over longer periods of time and to be more reciprocated" (Tsvetkova et al. 2017).

2. Recognizing this problem, Wikipedia exhorts its community to welcome newcomers, with uneven success "Wikipedia, please do not bite the newcomers." (Wikipedia 2024b)

3. This commentary uses the term "Internet services" to refer to social media, user-generated content (UGC) services, and online platforms.

10.54501/jots.v2i2.169

Striking this balance is one of the hardest challenges that Internet services must navigate, and yet the US regulatory policy currently lets services prioritize the best interests of their audiences over regulators' paranoia about bad faith actors. That regulatory deference is in constant jeopardy. Should it change, it will hurt the Internet—and all of us.

2 Why assuming good faith online has gotten harder over time

Until the early 1990s, much of the Internet was governed by the National Science Foundation (NSF) rule restricting commercial activity online.[4] This restriction implicitly limited who had Internet access. Most users were affiliated with educational institutions, government agencies, and the military—the primary entities allowed to connect to the Internet under the NSF rules. Due to those affiliations, most users were either employed or in school, so they were a more highly educated community compared to the general population.[5]

Furthermore, the Internet lacked user-friendly client-side software, so Internet users needed some technological sophistication: "The pre-Web Internet was an almost entirely text-based world...[including] command line driven programs such as Archie, which we used to try to find particular files. If this makes the pre-Web sound like a place that was only welcoming to techies in those days, you're right, it was" (Vaughan-Nichols 2011).[6]

Because of these dynamics, Internet users in the early 1990s were fairly homogeneous (Wong 2021): mostly male,[7] technologically savvy,[8] affluent (unless they were students), and educated (Wong 2021).

This homogeneity, though undesirable for many reasons, had an unexpected benefit: it created an environment where assuming good faith by other users was not wholly irrational. To the extent that service designers and users shared demographic attributes, the designers were more likely to anticipate and discourage obvious misbehavior from users who were like them.[9] Furthermore, homogeneous users are more likely to share the same biases and predilections, so their behavior would feel normal to each other even if it would have excluded or harmed more diverse users.

Two other characteristics of the early Internet further contributed to an environment where users' good faith could be assumed. First, the online population was much smaller, which made misbehavior less compelling because less money and fame was at stake (Internet World Stats 2023). Also, the smaller population increased the odds that people would personally know and have repeat interactions with each other, which helped more

4. The NSF funded the NSFNET, the Internet's principal backbone at the time, and its funding was restricted primarily to "research and education in the sciences and engineering." Accordingly, the NSFNET Acceptable Use Policy prohibited (among other things) "use for for-profit activities" and "extensive use for private or personal business" (NSF 1993; Cerf 2019; Leiner et al. 2009).

5. In a 1994 survey of web users, "45% of the respondents describe themselves as professionals, and 22% as graduate students" (Pitkow and Recker 1994).

6. For another perspective, see Wheaton (2022).

7. In a 1994 survey of web users, 94% reported as male (Pitkow and Recker 1994). "Group discourse [in 1982] reflected the leisure pursuits of young male engineers and computer scientists—science fiction, football, ham radios, cars, chess, and bridge" (Rosenzweig 1998). For demographic statistics for computer ownership in 1994 and 1997, a rough proxy for who had Internet access, see NTIA (1998).

8. In a 1994 survey of web users, "Most people (77%) had over ten years of programming experience and knew six to ten programming languages (41%)" (Pitkow and Recker 1994).

9. "[D]esigners and managers often assume their users are 'just like us'" (quoting a content moderation manager). Site designers "think of their own usage of social media and their friends' usage, and design their policies on the presumption that the site will be used by people in good faith who have the same definitions that they do as to what's unacceptable" (Gillespie 2018).

disputes resolve informally.[10] Second, while anonymous online conduct was possible, a lot of online activity was attributed or attributable to users.[11] Indeed, many users adhered to community norms to build their reputational capital in the community. Though the early Internet was not idyllic, early Internet users could often assume good faith by other Internet users.

Over the past three decades, the Internet has changed in many respects, including population demographics and size (Pew Research Center 2021; Georgia Tech GVU Center 1998). Now, online communities routinely cater to millions or even billions of users simultaneously—and those users geographically span the globe and demographically span the spectrum on every characteristic. As services scale up, they can no longer rely on social and demographic similarities between community members: "It is easier to maintain any given norm in a smaller community than a larger one. As a community grows, it becomes easier for individuals and groups to resist a norm" (Grimmelmann 2017). In other words, good-faith assumptions do not scale. Instead, the Internet's expansion and evolution has made it improbable, if not impossible, to assume other users' good faith.

The "Eternal September" illustrated how the Internet's demographic shift posed challenges for the prevailing good-faith presumptions. Starting in September 1993 (Grossman 1998), some commercial online services, including AOL, allowed users to access USENET, an early cross-platform online message board service (Pfaffenberger 1994; Gregory, Mann, and Parker 1995; Harrison 1995; Spencer and Lawrence 1998; Rittner 1997). This interconnection unleashed a flood of new users on USENET whose demographics and technological sophistication differed from existing users and who did not adhere to prevailing community norms. A similar dynamic occurred as incoming college freshmen in August and September could access the Internet for the first time without fully understanding the online communities they were engaging with. As the Internet later became even more widely accessible, it experienced an endless stream of new users—hence the term "Eternal September." The existing Internet community did not eagerly welcome these newcomers (Smith 2020). Instead, veteran users derided newcomers and assumed their bad faith instead of good. This implicit conflict between incumbents and newcomers has played out countless times online—on individual services and across the Internet generally—due to tribalism, xenophobia, and cultural differences. Any effort to assume other users' good faith usually becomes collateral damage.

Today, "clueless noobs" are not the biggest threat to the modern Internet's norms.[12] Instead, determined malefactors, including cybercriminals and state-sponsored attackers, routinely target Internet services.[13] With spammers, trolls, and jerks added to the scene, it has become functionally impossible today for online communities to assume that all users come to them in good faith. Instead, bad-faith actors attack every service and pose immediate and substantial threats to a community's integrity. In one notorious example, the *Los Angeles Times* shut down a wiki feature within 48 hours of launch because vandals had already overrun it (Rainey 2005; Grimmelmann 2017). Such threats must be quickly vanquished before they corrode the community (Kumar et al. 2018). Accordingly, new Internet services must assume that bad-faith actors are coming for them.

10. Among a small number of repeat players, community norms can mitigate disputes among community members (Ellickson 1991). The Internet's relatively small size also fostered visions that the Internet could self-manage: "Where there are real conflicts, where there are wrongs, we will identify them and address them by our means" (Barlow 1996).

11. In one famous early example, astronomer Cliff Stoll unmasked an anonymous cybercriminal (Stoll 1989).

12. Grimmelmann taxonomized the top concerns as "congestion, cacophony, abuse, and manipulation" (Grimmelmann 2017, 53–55).

13. "These spaces have been infiltrated by malicious state actors and self-identified insurrectionists. They use the same trolling techniques…to undermine our institutions, our communities and our trust in one another, in known facts and in our democracy" (Wong 2021).

Despite the overwhelming evidence of this phenomenon, even today some entrepreneurs still embrace a romanticized view of how their users will behave. As one content moderator explained, "Everybody wants their site to be a place where only Good Things happen, and when someone is starting up a new user-generated content site, they have a lot of enthusiasm and, usually, a lot of naïveté" (Gillespie 2018). For example, over the past few years, new services such as Gab, Rumble, Parler, Gettr, and Truth Social have positioned themselves as "free speech" alternatives to the incumbent services. To the extent they do less content moderation than the incumbents, the results have been unsurprising.[14] These services "speedrun" (Masnick 2020, 2021) through content moderation techniques, and quickly add policies and procedures that they should have adopted pre-launch, as they scramble to address the multitudinous threats facing their community (Binder 2022; Krishan 2022; Petrizzo 2022). The new entrants also have learned firsthand that malefactors will exploit all of the attack surfaces (Greenberg 2021; Binder 2021).

Thus, the appropriateness of assuming user good faith online has faded over the past three decades. As former Google and Twitter (now "X") lawyer Nicole Wong explained, in the 2000s, she "did not foresee the broad and coordinated weaponization of these open and free spaces that we built and advocated for. For a period, the bad actors could be managed or minimized. But, over time, these spaces have become playgrounds for trolls" (Wong 2021). This evolution reflects the Internet's loss of innocence more generally.

Today, presumptions of user good faith are—at best—quaintly anachronistic. As LinkedIn cofounder Reid Hoffman said, "Wild idealism was the lingua franca of Web 2.0" (Hoffman 2021). Now, failing to anticipate and prepare for malefactor attacks could be considered trust-and-safety malpractice.[15]

3 How services can anticipate bad-faith users

The historical preconditions that allowed services to assume users' good faith online have been overtaken by the inevitability that bad-faith user conduct will occur. Three ideas about how services seeking to foster users' pro-social activities can anticipate and mitigate pernicious users.

3.1 Adversarial wargaming and Trust-and-Safety by design

An Internet service can never completely predict users' behavior.[16] This unpredictability sometimes leads to positive and generative innovations, such as Twitter's retweet function, which replaced users' manual attempts to retweet (Kantrowitz 2019). Nevertheless, Internet services should identify and plan for the foreseeable and inevitable unwanted behavior.

For example, many services—including Craigslist, Facebook, and YouTube—allow users to report or "flag" problematic content from other users (Craigslist, n.d.; Meta, n.d.; YouTube, n.d.). This kind of user-driven feedback sounds helpful in theory, but users will coordinate their actions to submit false reports on legitimate content for improper purposes, a

14. "In the early days of the Web 2.0 era, we may have aspired to the wisdom of the crowd. But the way things played out, we often simply got the madness of the masses" (Hoffman 2021).

15. "Entrepreneurs, designers, and technologists building digital platforms that significantly impact the lives of billions of people…[must] actively increas[e] our efforts to put checks on bad actors and lowest-common-denominator impulses" (Hoffman 2021). See also Rigot (2022).

16. For example, child-friendly services could not prevent unwanted misbehavior even when they tightly structured users' ability to talk with each other (Copia Institute 2020).

phenomenon called "brigading" (Merriam-Webster 2023). If a service launches a user reporting tool without designing the system to thwart brigading, it may be doing more harm than good.

Internet services can better anticipate less common bad-faith misuses by engaging in adversarial wargaming (i.e., before launch, brainstorm scenarios from the perspective of bad-faith actors and stress-test the tools accordingly).[17] Stress tests need to be done early enough that the problems can be fixed before launch, and product managers (and their executives) need to take any concerns seriously even when risk probabilities are low or when only a small subset of the user population will be affected.

Some design weaknesses may become apparent during beta tests, and services can retain third-party white-hat consultants to provide additional testing. However, the most valuable insights will come from the service's in-house trust & safety and content review teams. After all, those teams will have to deal with any problems after launch, and they have firsthand knowledge of the specific ways that bad-faith actors are already misusing the service.

Thus, services should embrace a "trust-and-safety by design" approach, analogous to the "privacy by design" principle (European Union 2016; Cavoukian 2011). The in-house trust-and-safety and content review teams should play a key role early in the development of the service's specifications. These teams can do their own adversarial wargaming and red teaming (Botsman 2022). Respecting the feedback from these internal experts will reduce the number and severity of malefactor-caused problems post-launch.

3.2 Design the service to encourage pro-social behavior

Internet services can channel users toward pro-social behavior and away from bad-faith misuse based on how they design their offerings (Douek 2022; Grimmelmann 2017; Katyal 2003).[18] Nirav Tolia, who cofounded Epinions (a consumer review service) and Nextdoor (a local social network), described three levers that Internet services can use to shape user behavior (Tolia 2021).

The first lever is the service's "structure," including the prompts for user submissions and the forms used to capture those submissions (Grimmelmann 2017). As an example of structure, Tolia noted how Twitter limits the number of characters in a tweet (Tolia 2021). Another example might be Nextdoor's "kindness reminder," which "reminds the member about Nextdoor's Community Guidelines, and gives that member time to reflect, and hopefully refrain from posting a comment that doesn't comply with our Guidelines" (Nextdoor, n.d.).

The second lever is the "incentive," which is a user's motivation for submitting content, and the third is user "reputation," including the attributability of a user's content and behavior (Farmer and Glass 2010). Every design choice sends important signals to users, so services need to develop a mixture of structure, incentive, and reputation that is likely to elicit the kind of user activities the service desires (Tolia 2021).

These three levers are interconnected. Take, for example, a service that pays users for submitting content. Such incentives may motivate good-faith actors to submit for the wrong reasons and will also inevitably attract bad-faith users seeking to maximize

17. Some services already deploy in-house adversarial expertise. For example, Niantic employs an "Adversarial Planning Lead, Trust & Safety" with responsibilities to "design and conduct threat assessments with our game and product teams, lead red team exercises and other threat ideation work to ensure we address potential harms to our users" (Niantic Labs 2022).

18. Services can design features that "nudge" people to make better decisions (Thaler and Sunstein 2009). Hoffman (2021) described how Internet services can "leverag[e] humanity's less virtuous impulses." As Nirav Tolia said, "We want to nudge people to be good" (Tolia 2021).

payouts without providing the desired content (Tolia 2021). If the service wants to use this incentive, it will need to design the payouts anticipating all of these dynamics. The service can also use other levers to discourage bad-faith submissions by restricting payouts only to users who have a good reputation or by adding barriers to the content submission process to reduce the profitability of illegitimate submissions.[19]

If a service can optimize the mix of levers, users acting in their own self-interest will naturally take actions that enhance the community. However, it is extremely unlikely that a service will set the configuration perfectly on day one. Every service will inevitably tinker with the levers over time, based on new insights and conditions.[20]

3.3 Diversify the team

Homogeneous development teams have significant blind spots. They will fail to anticipate otherwise-foreseeable bad-faith uses, as well as ways the service may be unintentionally harming or disadvantaging user subpopulations lacking representation among the developer team. Increasing the development team's diversity and taking their perspectives seriously during the development process reduces the service's blind spots.[21] Diversifying the development team is essential for an Internet service's success. It is also the right thing to do.

4 Policy implications

This commentary, so far, has considered how services and users interact, without regard to the legal backdrop. For the most part, that is because many design choices are—and should be—driven by a service's business objectives, not legal concerns. Ultimately, everyone benefits when Internet services can determine the best solutions for their communities (Goldman 2019a).

The current design freedoms enjoyed by services are coming under extraordinary pressure. Regulators are increasingly requiring Internet services to reduce the risks associated with publishing third-party content. For example, the UK Online Safety Act imposes a duty of care on Internet services to curb antisocial behavior (UK 2023; Goldman 2019b), where regulators can point to every antisocial incident as prima facie evidence that the services insufficiently satisfied their duty. Similarly, the European Union's Digital Services Act requires some Internet services to mitigate systemic risks arising from user-generated content (EU 2022).

Intentionally or not, these legal duties unravel services' ability to assume any good-faith actions by users.[22] When regulations hold services accountable for the bad-faith activity—which inevitably occurs on every service—Internet services will view every user as a potential malefactor who can create ruinous liability for them. To manage these risks, the services will harden their systems—their structure, incentives, reputation, and content

19. For example, Epinions required user-submitted reviews to exceed a minimum word count, which thwarted bad-faith users who sought the payouts for low amounts of effort (Tolia 2021).

20. Hoffman (2021) writes, "Even if you somehow manage to get it right the first time, things will eventually change in ways that make additional adaptation necessary." Similarly, Botsman (2022) says, "As unanticipated consequences become apparent, it's up to entrepreneurs to implement, upgrade, or completely rethink the business models and structural mechanisms they have in place to reduce the negative impacts."

21. "We need leaders with empathy for people who are experiencing harassment. We need people who are from the groups that keep getting pushed off platforms—the Black and Latinx, Indigenous, and Asian users, women and/or nonbinary users, transgender users, and disabled users" (Pao 2021). See also Rigot (2022).

22. "I fear that there is too much focus on the bad actors, and we're so busy trying to remove or eliminate or punish the bad actors that we're not spending enough time trying to figure out ways to encourage and amplify and bring along the good ones" (Tolia 2021).

moderation functions—against all actual and potential threats. Hardened systems do not foster environments that attract and encourage good-faith users. Instead, good-faith users will be deterred or driven off by the less inviting environment, and because Internet services cannot avoid bad-faith activity no matter how hardened they are, the services will migrate away from allowing user-generated content entirely (Goldman 2019b).

In contrast, the current US policy, codified in 47 U.S.C. § 230 (1996), enables Internet services to strike more productive balances. Section 230 allows Internet services to assume users' good faith without imposing liability for the inevitable bad faith actions that will follow (Goldman 2019b). Further, Section 230 provides Internet services with the legal freedom to experiment with different site designs and configurations to combat bad-faith actions without fearing liability for any omissions or for tacitly admitting that prior solutions did not work. As Tolia (2021) said, "Let's focus on the positive and let's build systems that reinforce that." If we want services to keep assuming users' good faith and catering to good-faith actors, Section 230 provides the legal foundation that preserves that possibility.

References

Ayers, Phoebe, Charles Matthews, and Ben Yates. 2008. *How Wikipedia Works: And How You Can Be a Part of It.* No Starch Press. ISBN: 978-1593271763.

Barlow, John Perry. 1996. "A Declaration of the Independence of Cyberspace." *Electronic Frontier Foundation* (February 8, 1996). https://www.eff.org/cyberspace-independence.

Binder, Matt. 2021. "GETTR, the Newest Pro-Trump Social Network, Was Hacked on Launch Day and Is Now Fighting with Furries." *Mashable* (July 6, 2021). https://mashable.com/article/gettr-hacked.

———. 2022. "Truth Social Already Censoring Content, Bans User Who Made Fun of Trump Media CEO." *Mashable* (February 22, 2022). https://mashable.com/article/trump-truth-social-free-speech-bans.

Botsman, Rachel. 2022. "Tech Leaders Can Do More to Avoid Unintended Consequences." *Wired* (May 24, 2022). https://www.wired.com/story/technology-unintended-consequences/.

Bregman, Rutger. 2020. *Humankind: A Hopeful History.* Bloomsbury Publishing. ISBN: 9780316418553.

Cavoukian, Ann. 2011. *Privacy by Design: The 7 Foundational Principles.* Research report. Information and Privacy Commissioner of Ontario. https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf.

Cerf, Vinton G. 2019. "In debt to the NSF." *Communications of the ACM* 62, no. 4 (March): 5–5. ISSN: 1557-7317. https://doi.org/10.1145/3313989.

Copia Institute. 2020. *Creating Family Friendly Chat More Difficult than Imagined (1996).* Research report. Trust and Safety Foundation. https://trustandsafetyfoundation.org/blog/creating-family-friendly-chat-more-difficult-than-imagined-1996/.

Craigslist. n.d. "Flags and Community Moderation." Accessed January 17, 2024. https://www.craigslist.org/about/help/flags_and_community_moderation.

Douek, Evelyn. 2022. "Content moderation as systems thinking." *Harvard Law Review* 136:526. https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/.

Ellickson, Robert C. 1991. *Order Without Law: How Neighbors Settle Disputes.* Harvard University Press. ISBN: 978-0674641693.

European Union. 2016. "Article 25, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)." https://eur-lex.europa.eu/eli/reg/2016/679/oj.

———. 2022. "Article 35, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)." https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A32022R2065.

Farmer, Randy, and Bryce Glass. 2010. *Building Web Reputation Systems.* O'Reilly Media, Inc. ISBN: 978-0596159795.

Georgia Tech GVU Center. 1998. "GVU's 10th WWW User Survey." https://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press. ISBN: 9780300173130.

Goldman, Eric. 2010. "Wikipedia's labor squeeze and its consequences." *Journal of Telecommunications and High Technology Law* 8:157.

———. 2019a. "Internet immunity and the freedom to code." *Communications of the ACM* 62, no. 9 (August): 22–24. ISSN: 1557-7317. https://doi.org/10.1145/3349270.

———. 2019b. "The U.K. Online Harms White Paper and the Internet's Cable-ized Future." *SSRN Electronic Journal,* ISSN: 1556-5068. https://doi.org/10.2139/ssrn.3438530.

Greenberg, Andy. 2021. "An Absurdly Basic Bug Let Anyone Grab All of Parler's Data." *Wired* (January 12, 2021). https://www.wired.com/story/parler-hack-data-public-posts-images-video/.

Gregory, Kate, Jim Mann, and Tim Parker. 1995. *Using Usenet Newsgroups.* Que Corp. Publishing. ISBN: 9780789701343.

Grimmelmann, James. 2017. "The Virtues of Moderation." *Yale Journal of Law and Technology* (May). https://doi.org/10.31228/osf.io/qwxf5.

Grossman, Wendy. 1998. *Net.wars.* NYU Press. ISBN: 9780814731062.

Harrison, Mark. 1995. *The USENET Handbook: A User's Guide to Netnews.* O'Reilly Media. ISBN: 9781565921016.

Hoffman, Reid. 2021. "Human Nature in Vices and Virtues: An Adam Smith Approach to Building Internet Ecosystems and Communities." https://knightfoundation.org/human-nature-in-vices-and-virtues-an-adam-smith-approach-to-building-Internet-ecosystems-and-communities/.

Internet World Stats. 2023. "Internet Growth Statistics." https://www.Internetworldstats.com/emarketing.htm.

Kantrowitz, Alex. 2019. "The Man Who Built the Retweet: 'We Handed A Loaded Weapon To 4-Year-Olds'." *Buzzfeed* (July 23, 2019). https://www.buzzfeednews.com/article/alexkantrowitz/how-the-retweet-ruined-the-Internet.

Katyal, Neal Kumar. 2003. "Digital Architecture as Crime Control." *The Yale Law Journal* 112, no. 8 (June): 2261. ISSN: 0044-0094. https://doi.org/10.2307/3657476.

Krishan, Nihal. 2022. "Truth Social Criticized by Far-Right Talk Show Host for 'Censorship' as It Surges in Popularity." *Washington Examiner* (February 25, 2022). https://www.washingtonexaminer.com/policy/truth-social-faces-conservative-criticism-for-censorship-as-it-surges-in-popularity.

Kumar, Srijan, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. "Community interaction and conflict on the web." In *Proceedings of the 2018 world wide web conference,* 933–43. https://doi.org/10.1145/3178876.3186141.

Leiner, Barry M., Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. 2009. "A brief history of the Internet." *ACM SIGCOMM Computer Communication Review* 39 (5): 22–31. https://doi.org/10.1145/1629607.1629613.

Masnick, Mike. 2020. "Parler Speedruns the Content Moderation Learning Curve; Goes from 'We Allow Everything' to 'We're the Good Censors' in Days." *Techdirt* (July 1, 2020). https://www.techdirt.com/articles/20200630/23525844821/parler-speedruns-content-moderation-learning-curve-goes-we-allow-everything-to-were-good-censors-days.shtml.

———. 2021. "Trumpist Gettr Social Network Continues to Speed Run Content Moderation Learning Curve: Bans, Then Unbans, Roger Stone." *Techdirt* (August 26, 2021). https://www.techdirt.com/articles/20210825/17204647438/trumpist-gettr-social-network-continues-to-speed-run-content-moderation-learning-curve-bans-then-unbans-roger-stone.shtml.

Merriam-Webster. 2023. "Calling In a New 'Brigade'." https://www.merriam-webster.com/words-at-play/brigading-online-poll-meaning.

Meta. n.d. "How Do I Report Inappropriate or Abusive Things on Facebook." Accessed January 17, 2024. https://www.facebook.com/help/212722115425932.

Nasaw, Daniel. 2012. "Meet the 'Bots' That Edit Wikipedia." *BBC* (July 25, 2012). https://www.bbc.com/news/magazine-18892510.

National Telecommunications and Information Administration. 1998. *Falling through the Net II: New Data on the Digital Divide.* Research report. NTIA. http://www.ntia.doc.gov/ntiahome/net2/.

Nextdoor. n.d. "About the Kindness Reminder." Accessed January 17, 2024. https://help.nextdoor.com/s/article/About-the-Kindness-Reminder.

Niantic Labs. 2022. "Niantic Careers: Openings." https://nianticlabs.com/careers/openings?hl=en.

Office of Inspector General, National Science Foundation. 1993. *Review of NSFNET.* Research report. NSF. https://www.nsf.gov/pubs/stis1993/oig9301/oig9301.txt.

Pao, Ellen. 2021. "Knowing What You Know Now about the Internet and How Your Venture Turned Out." https://knightfoundation.org/ellen-pao.

Petrizzo, Zachary. 2022. "Can You Call Pro-'Free Speech' Gettr's Billionaire Backer a 'Spy' on the App? We Tested It." *Yahoo News* (February 17, 2022). https://news.yahoo.com/call-free-speech-gettr-billionaire-191021503.html.

Pew Research Center. 2021. "Internet/Broadband Fact Sheet." https://www.pewresearch.org/Internet/fact-sheet/Internet-broadband/.

Pfaffenberger, Bryan. 1994. *The USENET Book: Finding, Using, and Surviving Newsgroups on the Internet.* Addison-Wesley Longman Publishing Co., Inc. ISBN: 9780201409789.

Pitkow, James, and Mimi Recker. 1994. "Results from the first world-wide web user survey." *Computer Networks and ISDN Systems* 27 (2): 243–54. https://doi.org/10.1016/0169-7552(94)90138-4.

Rainey, James. 2005. "'Wikitorial' Pulled Due to Vandalism." *Los Angeles Times* (June 21, 2005). https://www.latimes.com/archives/la-xpm-2005-jun-21-na-wiki21-story.html.

Reagle, Joseph Michael. 2010. *Good Faith Collaboration: The Culture of Wikipedia.* MIT Press. ISBN: 978-0262014472.

Rigot, Afsaneh. 2022. “If Tech Fails to Design for the Most Vulnerable, It Fails Us All.” *Wired* (May 15, 2022). https://www.wired.com/story/technology-design-marginalized-communities/.

Rittner, Don. 1997. *Rittner’s Field Guide to Usenet.* MNS Publishing. ISBN: 9780937666500.

Rosenzweig, Roy. 1998. “Wizards, bureaucrats, warriors, and hackers: Writing the history of the Internet.” *The American Historical Review* 103 (5): 1530–52. https://doi.org/10.1086/ahr/103.5.1530.

Smith, Ernie. 2020. “No More Eternal Septembers.” *Tedium* (October 13, 2020). https://tedium.co/2020/10/13/eternal-september-modern-impact/.

Spencer, Henry, and David Lawrence. 1998. *Managing Usenet.* O’Reilly & Associates, Inc. ISBN: 9781565921986.

Stoll, Clifford. 1989. *The Cuckoo’s Egg: Tracking a Spy through the Maze of Computer Espionage.* Pocket Books. ISBN: 0385249462.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Penguin. ISBN: 9780141040011.

Tolia, Nirav. 2021. “Interview With Eric Goldman,” October 29, 2021. https://knightfoundation.org/interview-nirav-tolia-with-eric-goldman/.

Tsvetkova, Milena, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri. 2017. “Even good bots fight: The case of Wikipedia.” *Public Library of Science* 12 (2): e0171774. https://doi.org/10.1371/journal.pone.0171774.

United Kingdom, U.K. Online Safety Act (2023), https://www.legislation.gov.uk/ukpga/2023/50/enacted.

Protection for private blocking and screening of offensive material, Stat. (1996), https://www.law.cornell.edu/uscode/text/47/230.

Vaughan-Nichols, Steven. 2011. “Before the Web: the Internet in 1991.” *ZDNet* (April 17, 2011). https://www.zdnet.com/article/before-the-web-the-internet-in-1991/.

Wheaton, Wil. 2022. “The Internet Used to be Smaller and Nicer. Let’s Get It Back.” *Wall Street Journal* (June 3, 2022). https://www.wsj.com/articles/the-internet-used-to-be-smaller-and-nicer-lets-get-it-back-11654261200.

Wikipedia. 2023, s.v. “Bots.” Last edited July 16, 2023, 21:05. https://en.wikipedia.org/wiki/Wikipedia:Bots.

———. 2024a, s.v. “Assume good faith.” Last edited January 25, 2024, 03:17. https://en.wikipedia.org/wiki/Wikipedia:Assume_good_faith.

———. 2024b, s.v. “Please do not bite the newcomers.” Last edited January 17, 2024, 20:22. https://en.wikipedia.org/wiki/Wikipedia:Please_do_not_bite_the_newcomers.

Wong, Nicole. 2021. “Lessons from the First Internet Ages.” https://knightfoundation.org/nicole-wong.

YouTube. n.d. “Report Inappropriate Videos, Channels, and Other Content on YouTube.” Accessed January 17, 2024. https://support.google.com/youtube/answer/2802027.

Zheng, Lei, Christopher M. Albano, Neev M. Vora, Feng Mai, and Jeffrey V. Nickerson. 2019. “The roles bots play in Wikipedia.” *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–20. https://doi.org/10.1145/3359317.

Authors

Eric Goldman (egoldman@gmail.com) is a Professor of Law, Associate Dean for Research, and Co-Director of the High Tech Law Institute, Santa Clara University School of Law. He is a founding board member of the Trust & Safety Professional Association and the Trust & Safety Foundation. Website: http://www.ericgoldman.org.

Acknowledgements

This commentary was sponsored by a grant from the Center for Growth & Opportunity at Utah State University. Eric thanks Jeff Lazarus, Kaofeng Lee, Tsu Li Liew, and Jess Miers for their comments.

Keywords

Assume good faith; content moderation; Internet history; Section 230; trust and safety; Wikipedia.