Assuming Good Faith Online

Author:

Eric Goldman[a]

November 2022

Working Paper

[a] Eric Goldman is a law professor as well as the Associate Dean for Research and Co-Director of the High Tech Law Institute, Santa Clara University School of Law.

Website: http://www.ericgoldman.org Email: egoldman@gmail.com

* Author's note: I was General Counsel for Epinions from 2000–02. In 2021, in my role as a Knight Visiting Scholar, I helped obtain some of the Knight Foundation essays and interviews I cite (as part of the *Lessons from the First Internet Ages* series). I thank Jeff Lazarus, Kaofeng Lee, Tsu Li Liew, and Jess Miers for their comments.

The Center for Growth and Opportunity at Utah State University

Introduction

One of Wikipedia's fundamental principles is to "assume good faith."[1] The Wikipedia project page explains that "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."[2] In theory, this principle should not be remarkable. Most people act in good faith most of the time—and any well-functioning society depends on good-faith interactions being the norm. Nevertheless, Wikipedia's assume-good-faith principle feels remarkable because it defies the widespread and obvious evidence of bad-faith activity by Internet users, including 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.[3]

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."[4] In the face of multitudinous threats, the Wikipedia community does not always wait for that obvious evidence. Instead, Wikipedia relies heavily on bots to patrol its premises and take action,[5] even though bots cannot "assume" good faith.[6] Furthermore, the Wikipedia community has developed xenophobic tendencies[7] that lead to skepticism of newcomers' activities.[8] Perhaps battle-scarred after two decades of fighting bad-faith activity, Wikipedia struggles to maintain 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 enabling user-generated content faces a similar dilemma of balancing good-faith and bad-faith activity.[9] Without that balance, the service loses 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 drives towards one of two suboptimal outcomes. Either it devolves into a cesspool of bad-faith activity or becomes a restrictive locked-down environment with limited expressive options for any user, even well-intentioned ones.

---

1 "Wikipedia: Assume Good Faith," Wikimedia Foundation, accessed February 12, 2022, https://en.wikipedia.org/wiki/Wikipedia:Assume_good_faith.

2 Wikimedia, "Assume Good Faith;" see also Joseph Reagle, *Good Faith Collaboration: The Culture of Wikipedia* (Cambridge, MA: MIT Press, 2010), 60–64; Phoebe Ayres, Charles Matthews, and Ben Yates, *How Wikipedia Works and How You Can Be a Part of It* (San Francisco: No Starch Press, 2008), 366–67.

3 Eric Goldman, "Wikipedia's Labor Squeeze and Its Consequences," *Journal of Telecommunications and High Technology Law* 8, no. 1 (Winter 2010): 157–83.

4 Wikimedia, "Assume Good Faith."

5 "Bots are playing an increasingly important role in the creation of knowledge in Wikipedia." Lei Zheng, Christopher M. Albano, Neev M. Vora, Feng Mai, and Jeffrey V. Nickerson, "The Roles Bots Play in Wikipedia," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 2019), https://dl.acm.org/doi/pdf/10.1145/3359317. See also Daniel Nasaw, "Meet the 'Bots' That Edit Wikipedia," BBC, July 25, 2012, https://www.bbc.com/news/magazine-18892510; and "Wikipedia: Bots," Wikimedia Foundation, last modified June 2, 2022, https://en.wikipedia.org/wiki/Wikipedia:Bots.

6 "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." Milena Tsvetkova, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri, "Even Good Bots Fight: The Case of Wikipedia," *PLoS ONE* 12, no. 2 (2017), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171774.

7 Goldman, "Wikipedia's Labor Squeeze."

8 Recognizing this problem, Wikipedia aspires to be friendly toward newcomers. Arguably, this principle is honored more often in the breach. "Wikipedia: Please Do Not Bite the Newcomers," Wikimedia Foundation, accessed February 12, 2022, https://en.wikipedia.org/wiki/Wikipedia:Please_do_not_bite_the_newcomers.

9 These services go by many names, including "social media," UGC services, and "platforms." This essay refers to them as "Internet services."

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

Why Assuming Good Faith Has Gotten Harder over Time

Until the early 1990s, much of the Internet was still governed by the National Science Foundation (NSF) rule restricting commercial activity online.[10] This restriction implicitly limited Internet access. Most users were affiliated with educational institutions, government agencies, and the military, as those were 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 more highly educated as a group when compared to the general population.[11] Furthermore, the Internet lacked user-friendly client-side software, so Internet users needed some technological savvy.[12]

In practice, this meant Internet users in the early 1990s were fairly homogeneous:[13] mostly male,[14] technologically savvy,[15] affluent (unless they were students), and educated.[16] 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

10 The NSF funded the NSFNET, the Internet's principal backbone at the time, and its funding was restricted to being used "primarily for 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." *Review of NSFNET*, Office of Inspector General, National Science Foundation, March 23, 1993, https://www.nsf.gov/pubs/stis1993/oig9301/oig9301.txt. Also, regarding the NSF's role in the Internet's development, see Vinton G. Cerf, "In Debt to the NSF," *Communications of the ACM* 62, no. 4 (April 2019): 5; and Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff, "A Brief History of the Internet," *ACM SIGCOMM Computer Communication Review* 39, no. 5 (October 2009): 22.

11 In a 1994 survey of web users, "45% of the respondents describe themselves as professionals, and 22% as graduate students." James E. Pitkow and Margaret M. Recker, "Results from the First World-Wide Web User Survey," *Journal of Computer Networks and ISDN Systems* 27, no. 2 (November 1994): 243–54, https://smartech.gatech.edu/bitstream/handle/1853/3592/94-19.pdf.

12 "The pre-web Internet was an almost entirely text-based world….more typical though were 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." Steven Vaughan-Nichols, "Before the web: the Internet in 1991," ZDNet, April 17, 2011, https://www.zdnet.com/article/before-the-web-the-Internet-in-1991/; compare this to Wil Wheaton, "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.

13 Regarding the homogeneity of the early Internet developers, see Nicole Wong, "Lessons from the First Internet Ages," Knight Foundation, October 29, 2021, https://knightfoundation.org/nicole-wong.

14 In a 1994 survey of web users, 94 percent reported as male. Pitkow and Recker, "User Survey." For demographic statistics for computer ownership in 1994 and 1997, which is a rough but imperfect proxy for Internet access, see National Telecommunications and Information Administration, *Falling through the Net II: New Data on the Digital Divide*, July 28, 1998, http://www.ntia.doc.gov/ntiahome/net2/. "Group discourse [in 1982] reflected the leisure pursuits of young male engineers and computer scientists—science fiction, football, ham radios, cars, chess, and bridge." Roy Rosenzweig, "Wizards, Bureaucrats, Warriors, and Hackers: Writing the History of the Internet," *American History Review* 103, no. 5 (December 1998): 1530–52.

15 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, "User Survey."

16 The early Internet was "largely populated by the more educated, more white and more male segments of those societies." Whether or not they had access to "the Internet," users could access paywalled walled-garden commercial online services (such as CompuServe and Prodigy) and BBSes. These communities were also fairly homogeneous due to the technological sophistication required to access them plus the high costs of connecting to the service and acquiring the necessary equipment like the computer and modem. Wong, "Lessons."

discourage obvious misbehavior from users who were like them.[17] 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.[18] Also, the smaller population increased the odds that people would personally know each other and have repeat interactions with each other, which also helped more disputes resolve informally.[19] Second, while anonymous online conduct was possible, a lot of online activity was attributed to users.[20] Indeed, many users intentionally adhered to community norms in order to build their reputational capital in the community. The early Internet was not idyllic, but there were many reasons why early Internet users logically assumed good faith by other Internet users.

Over the past three decades, the Internet has changed in many respects, including population demographics and size.[21] 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 close social and demographic ties between community members because good-faith assumptions do not scale proportionally. Instead, the Internet's expansion and evolution has made it improbable, if not impossible, to assume good faith.

The "Eternal September" provided an early example of how a shift in demographics changed the Internet's character. Starting in September 1993,[22] some commercial online services, including AOL, allowed users to access USENET.[23] This interconnection unleashed a flood of new users whose demographics and technological sophistication differed from existing users and who did not adhere to prevailing community norms. A similar dynamic had already been at play each September as incoming college freshmen were given Internet access 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

---

17 "[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." Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (New Haven: Yale University Press, 2018), 14, 17.

18 "Internet Growth Statistics," Internet World Stats, last modified June 9, 2022, https://www.Internetworldstats.com/emarketing.htm.

19 "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." James Grimmelmann, "The Virtues of Moderation," Yale Journal of Law and Technology 17, no. 1 (2015): 42, 73. See also Robert C. Ellickson, *Order Without Law: How Neighbors Settle Disputes* (Cambridge, MA: Harvard University Press, 1994). "Where there are real conflicts, where there are wrongs, we will identify them and address them by our means." John Perry Barlow, "A Declaration of the Independence of Cyberspace," Electronic Frontier Foundation, February 8, 1996, https://www.eff.org/cyberspace-independence.

20 Clifford Stoll, *The Cuckoo's Egg: Tracking a Spy through the Maze of Computer Espionage* (New York: Pocket Books, 1989).

21 "Internet/Broadband Fact Sheet," Pew Research Center, April 7, 2021, https://www.pewresearch.org/Internet/fact-sheet/Internet-broadband/. For a snapshot from the late 1990s, see "GVU's 10th WWW User Survey," Georgia Tech GVU Center, 1998, https://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/.

22 Wendy M. Grossman, *Net.Wars* (New York: NYU Press, 1998).

23 USENET was an early cross-platform online message board service. See, for example, Bryan Pfaffenberger, *The USENET Book: Finding, Using, and Surviving Newsgroups on the Internet* (Boston: Addison-Wesley, 1994); Kate Gregory, Jim Mann, and Tim Parker, *Using Usenet Newsgroups* (Seattle: Que Corp. Publishing, 1995); Mark Harrison, *The USENET Handbook* (Sebastopol, CA: O'Reilly Media, 1995); Henry Spencer and David Lawrence, *Managing Usenet* (Sebastopol, CA: O'Reilly Media, 1998); Don Rittner, *Rittner's Field Guide to Usenet* (self-published?, 1997).

"Eternal September." The existing Internet community did not extend a welcome greeting to these newcomers.[24] Instead, veteran users derided newcomers, essentially assuming their bad faith instead of good. This incumbent versus newcomer dynamic has played out countless times online—on individual services and across the Internet generally—due to tribalism, xenophobia, and culture differences. The assumption of good faith usually becomes collateral damage.

Today, "clueless noobs" are not the biggest threat to the modern Internet.[25] Instead, determined actors, such as cybercriminals and state-sponsored attackers, routinely target Internet services.[26] Combine these malefactors with the standard complement of spammers, trolls, and jerks, and together they make it 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.[27] Such threats must be quickly vanquished before they fatally poison the community.[28] Accordingly, new Internet services must assume that bad-faith actors are coming for them.

Despite the overwhelming evidence of this phenomenon, 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."[29] For example, over the past few years, some new services have positioned themselves as "free speech" alternatives to the incumbent services.[30] To the extent they aspire to do less content moderation than the incumbents, the results have been unsurprising.[31] These services "speedrun" through new iterations of ways to moderate content[32] (i.e., quickly adding content moderation policies and procedures that they should have adopted pre-launch) when they realize that they must address the

24 Ernie Smith, "No More Eternal Septembers," Tedium, October 13, 2020, https://tedium.co/2020/10/13/eternal-september-modern-impact/.

25 For a different taxonomy of concerns, describing four online problems: "congestion, cacophony, abuse, and manipulation," see Grimmelmann, "Virtues," 53–55.

26 "These spaces have been infiltrated by malicious state actors and self-identified insurrectionists. They use the same trolling techniques, not just for entertainment, but to undermine our institutions, our communities and our trust in one another, in known facts and in our democracy." Wong, "Lessons."

27 One notorious example: the *Los Angeles Times* had to shut down a wiki feature within 48 hours of launch because it had already been overrun by vandals. James Rainey, "'Wikitorial' Pulled Due to Vandalism," *L.A. Times*, June 21, 2005, https://www.latimes.com/archives/la-xpm-2005-jun-21-na-wiki21-story.html. Also regarding the *L.A. Times* incident, see generally Grimmelmann, "Virtues."

28 See, for example, Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky, "Community Interaction and Conflict on the web," in *WWW 2018: Proceedings of the 2018 World Wide Web Conference* (Lyon, France: IW3C2, 2018), https://doi.org/10.1145/3178876.3186141.

29 Gillespie, "Custodians of the Internet," 17.

30 Examples include Gab, Rumble, Parler, Gettr, and Truth Social.

31 "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." Reid Hoffman, "Human Nature in Vices and Virtues: An Adam Smith Approach to Building Internet Ecosystems and Communities," Knight Foundation, October 29, 2021, https://knightfoundation.org/human-nature-in-vices-and-virtues-an-adam-smith-approach-to-building-Internet-ecosystems-and-communities/.

32 For example, Mike Masnick, "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; Mike Masnick, "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.

multitudinous threats to their community.[33] They also have learned first-hand that malefactors will attack them in other ways too.[34]

Thus, the appropriateness of assuming good faith has faded over the past three decades. As former Google and Twitter 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."[35] That is true of Wikipedia as well. Its "Assume Good Faith" policy is an idealistic holdover from Wikipedia's early days. As bad-faith activity online has increased, Wikipedia's assumption has become untenable. This evolution is a microcosm of the Internet's loss of innocence more generally. Today, any presumptions of good faith are—at best—quaintly anachronistic.[36] Instead, failing to prepare for malefactor attacks could be considered content moderation malpractice.[37]

How to Anticipate Bad-Faith Users

The historical conditions for assuming users' good faith online has been overtaken by the inevitability that bad-faith user conduct will occur. This necessitates a balancing act from services seeking to foster the pro-social activities of the many users who naturally engage in good-faith activity while simultaneously mitigating the pernicious users. Internet services can take numerous steps, including the following three ideas, to promote this balance.

Adversarial Wargaming and Content Moderation by Design

An Internet service initially plans for users to act in certain ways, but services cannot conjecture all of the ways that their users will actually use the service.[38] This unpredictability is not inherently bad; it can lead to positive and generative outcomes that improve the service.[39] Nevertheless, some users will misuse and attempt to game every feature offered by the service, and Internet services

33 For example, Matt Binder, "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; Nihal Krishan, "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; and Zachary Petrizzo, "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.

34 Andy Greenberg, "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/; Matt Binder, "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.

35 Wong, "Lessons."

36 "Wild idealism was the lingua franca of web 2.0." Hoffman, "Human Nature."

37 "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, "Human Nature"; Afsaneh Rigot, "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/.

38 "However hard I had anticipated it might be to convert emotions like pride and greed into more productive behaviors and aspirational identities, it has proven to be even harder in practice." Hoffman, "Human Nature." Compare this to Rachel Botsman, "Tech Leaders Can Do More to Avoid Unintended Consequences," *Wired*, May 24, 2022, https://www.wired.com/story/technology-unintended-consequences/.

39 For example, Twitter automated the retweet function to codify users' manual retweeting efforts. Alex Kantrowitz, "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.

should identify and plan for the threats it can foresee.[40] For example, many services give users a way to report or "flag" problematic content from other users.[41] This kind of user-driven feedback sounds helpful in theory; however, it is well-known that users will coordinate their actions to submit false reports on legitimate content for improper purposes, a pernicious phenomenon called "brigading."[42] If services do not design their systems to prevent brigading, their well-intentioned user reporting tool could actually harm the community.

Common bad-faith behavior like brigading may be easy to predict, but other potential problems can be identified only with careful planning and forethought. Internet services can better anticipate 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).[43] 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 feedback seriously even when risk probabilities are low or when only a small user population will be affected.

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

Thus, services should embrace "trust & safety by design" or "content moderation by design,"[44] analogous to the "privacy by design" principle,[45] whereby the in-house trust & safety and content moderation teams play a key role early in the development of the service's specifications and pre-launch adversarial wargaming.[46] Respecting the feedback from these teams of experts will reduce the number and severity of malefactor-caused problems post-launch.

---

40 However, even child-friendly services could not prevent unwanted misbehavior, no matter how tightly they structured the ability of users to talk with each other. Copia Institute, "Creating Family Friendly Chat More Difficult than Imagined (1996)," Trust and Safety Foundation, July 2020, https://trustandsafetyfoundation.org/blog/creating-family-friendly-chat-more-difficult-than-imagined-1996/.

41 For example, user flagging is an important content moderation tactic for Craigslist, Facebook, and YouTube. "Flags and Community Moderation," Craigslist, 2022, https://www.craigslist.org/about/help/flags_and_community_moderation; "How Do I Report Inappropriate or Abusive Things on Facebook," Meta, 2022, https://www.facebook.com/help/212722115425932; "Report Inappropriate Videos, Channels, and Other Content on YouTube," Google, 2022, https://support.google.com/youtube/answer/2802027.

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

43 Some services are already incorporating in-house adversarial perspectives. For example, Niantic has a job entitled "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." This job was advertised online earlier in 2022 at "Niantic Careers: Openings," Niantic Labs, https://careers.nianticlabs.com/openings/adversarial-planning-lead-trust-and-safety/.

44 For an argument in favor of ex ante approaches to content moderation, see Evelyn Douek, "Content Moderation as Administration," *Harvard Law Review* (forthcoming).

45 For example, Article 25 of the 2018 General Data Privacy Regulation, https://gdpr-info.eu/art-25-gdpr/; Ann Cavoukian, "Privacy by Design: The 7 Foundational Principles," Information and Privacy Commissioner of Ontario, January 2011, https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf.

46 Botsman, "Tech Leaders" (discussing how services should create their own internal "red teams").

Design the Service to Encourage Pro-Social Behavior

Internet services can channel users towards pro-social behavior and away from bad-faith misuse based on how they design their offerings.[47] Nirav Tolia, who co-founded Epinions (a consumer review service) and Nextdoor (a local social network), described three levers that Internet services can use to shape user behavior.[48] The first lever is the service's "structure," including the prompts for user submissions and the forms used to capture those submissions.[49] As an example of structure, Tolia noted how Twitter limits the number of characters in a tweet.[50] 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."[51] 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. 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.[52]

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 payouts without providing the desired content.[53] If the service wants to use this incentive, the service will need to design the payouts precisely. It can also use other levers to discourage the 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.[54]

If a service can optimize the mix of levers, users acting in their own self-interest will naturally take steps 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.[55]

---

47 See Grimmelmann, "Virtues"; and Neal Kumar Katyal, "Digital Architecture as Crime Control," *Yale Law Journal* 112, no. 8 (June 2003): 2261–89. Compare with the discussion of various ways that design features can "nudge" people to make better decisions in Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (New York: Penguin, 2008). As Nirav Tolia said, "We want to nudge people to be good." Nirav Tolia, "Lessons from the First Internet Ages," interview by Eric Goldman, Knight Foundation, October 29, 2021, https://knightfoundation.org/interview-nirav-tolia-with-eric-goldman/.

48 Tolia, "First Internet Ages." Compare with how Internet services can "leverag[e] humanity's less virtuous impulses" in Hoffman, "Human Nature." For a discussion of ways to design reputation systems, see Bryce Glass and Randy Farmer, *Building Web Reputation Systems* (Sebastopol, CA: O'Reilly Media, 2010).

49 Grimmelmann, "Virtues."

50 Tolia, "First Internet Ages." Hoffman gives several examples of how LinkedIn structured user behavior, such as preventing users from uploading headshots of themselves (to prevent their inevitable sexualization) and hiding the true number of connections over 500 to reduce connection-building for vanity purposes. Hoffman, "Human Nature."

51 "About the Kindness Reminder," Nextdoor Help Center, https://help.nextdoor.com/s/article/About-the-Kindness-Reminder.

52 Tolia, "First Internet Ages."

53 Tolia, "First Internet Ages."

54 For example, Epinions required user-submitted reviews to exceed a minimum word count. This thwarted bad-faith users who sought the payouts for low amounts of effort. Tolia, "First Internet Ages."

55 "Even if you somehow manage to get it right the first time, things will eventually change in ways that make additional adaptation necessary." Hoffman, "Human Nature." "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." Botsman, "Tech Leaders."

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 some user populations. Increasing the development team's diversity and taking their perspectives seriously during the development processes shrinks the service's blind spots.[56] Diversifying the development team is essential for an Internet service's success and is also the right thing to do.

Policy Implications

This essay, 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.[57]

The current design freedoms enjoyed by services are coming under extraordinary pressure. Regulators want Internet services to wave magic wands and precisely eliminate all bad-faith user activities. Premised on this assumption, regulators increasingly seek to compel Internet services to achieve this outcome.[58] Such fantastical dreams have the unwanted consequence of reducing or eliminating the services' ability to assume any good-faith actions by users.[59] If regulators require Internet services to eliminate all bad-faith activity, Internet services can do so only by starting with the presumption that every user is a malefactor and subsequently hardening their systems—their structure, incentives, reputation, and content moderation functions—to thwart every threat. This does not create environments that attract and encourage good-faith users. Instead, these designs lead to no-win outcomes for Internet services. Good-faith users will be driven off; and because Internet services cannot achieve the impossible, they will not be able to bear the associated legal risk of user-generated content.[60]

In contrast, the existing US policy, codified in Section 230, enables Internet services to achieve better balances.[61] Section 230 allows Internet services to assume users' good faith without imposing liability for the inevitable bad faith that will follow.[62] Further, Section 230 provides Internet

56 "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." Ellen Pao, "Knowing What You Know Now about the Internet and How Your Venture Turned Out," Knight Foundation, October 29, 2021, https://knightfoundation.org/ellen-pao/. See also Rigot, "If Tech Fails."

57 See Eric Goldman, "Internet Immunity and the Freedom to Code," *Communications of the ACM* 62, no. 9 (September 2019): 22–24.

58 For example, the UK Online Safety bill penalizes Internet services if they fail to eliminate anti-social behavior. Nominally, the bill regulates the services' efforts to create safe environments; but every anti-social incident provides prima facie evidence that the services' efforts were insufficient. Online Safety Bill, UK House of Commons Bill 285 (March 2022), https://publications.parliament.uk/pa/bills/cbill/58-02/0285/210285.pdf; Eric Goldman, "The UK Online Harms White Paper and the Internet's Cable-ized Future," *Ohio State Technology Law Journal* 16, no. 2 (Spring 2020): 351–62.

59 "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, "First Internet Ages."

60 Goldman, "UK Online Harms."

61 47 U.S.C. § 230.

62 Eric Goldman, "An Overview of the United States' Section 230 Internet Immunity," in *Oxford Handbook of Online Intermediary Liability*, ed. Giancarlo Frosio (Oxford, UK: Oxford University Press, 2020), 155.

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. 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.[63]

63 "Let's focus on the positive and let's build systems that reinforce that." Tolia, "First Internet Ages."