How to moderate an online discourse?

Parth Parikh
Department of Computer Science
North Carolina State University
Raleigh, NC 27606
Email: pmparikh@ncsu.edu

***Note*—I wrote this essay for my CSC-510 finals (Fall 2021).**

I. Where do we draw a line?

> Everything in moderation, including moderation.
> - Oscar Wilde

Is moderation the kind of censorship we agree with? Is censorship the kind of moderation we disagree with? [1] Is there a line? In my humble opinion, this is the main issue at stake.

In their article titled *The secret rules of the internet* [2], Catherine Buni and Soraya Chemaly argue that conclusively identifying why one content is acceptable but a slight variation breaks some policy remains the *holy grail of moderation*. They narrate an incident that the then senior content specialist at Youtube, Mora-Blanco, faced during the Green Movement. Her team had to address a gruesome video depicting the death of a young woman during the demonstrations against pro-government forces. As this was 2009, their guidelines lacked directives for ethical journalism involving graphic content. However, her team realizing the importance of the movement, decided to work with the legal department to modify Youtube's violence policy.

So, where exactly should we stand in this moderation spectrum? Should there be provisions for complete freedom? Constitution-mandated freedom? Centralized domain-specific freedom? Decentralized peer-reviewed freedom?

Regardless of the position, we take above, it is blatantly apparent that *civility* is the underlying goal of most discourse-based entities. In fact, *civility* is one of the founding principles of Hacker News (HN), a computer-science-focused news website. Paul Graham, the founder of HN, had the following to say on the welcome page [3]:

> Since long before the web, the anonymity of online conversation has lured people into being much ruder than they'd dare to be in person. So the principle here is not to say anything you wouldn't say face to face. This doesn't mean you can't disagree. But disagree without calling the other person names. If you're right, your argument will be more convincing without them.

In a way, the above principle discourages the practice of gratuitous negativity [4]. However, without any user action, these ideals dry out as mere ink on paper. Any platform relies on users to produce content, consume them, and moderate other users. For the latter, downvoting, vouching, and flagging can act as an effective virtual soapbox. So are the users solely to be blamed for any predicament? As rightly pointed out by an HN user [5], *the design of a medium is far more important than how it is moderated.* Over the last decade, with the advent of giants such as Facebook and Twitter, society has experienced the ramifications of designing recommendations around virality and engagement.

II. Moderation under a microscope

> The details of moderation practices are routinely treated as trade secrets.
> - Buni and Chemaly [2]

Is this problem philosophically NP? At the most fundamental level, given a discourse, a moderator has to analyze it and determine whether any moderation is absolutely necessary or not. While analyzing the notion of human beings solving real-life NP [1] problems, Scott Aaronson argued [6] that there could be no polynomial-time algorithm for recognizing such activities like great poetry or art. As the presence of such an algorithm would indicate that the task of composing them is in NP. Furthermore, he stated that humans could have the ability to answer special cases faster than a Turing machine. These could be search problems with a high degree of semantics, like proving Fermat's Last Theorem. Extending this discussion, one can argue that concretely determining the necessity of moderation is NP in nature, and therefore would require a degree of approximation for tractability.

Combining the above discussion with the one on user behavior, a natural question surfaces - what is a baseline approximation for any moderation? Well, constitution-mandated laws seem like a good starting point. A study by the European Parliament for their Digital Services Act had the following to say about the United States' regulatory and policy framework [7]:

> In the field of illegal content online of sexual character, the mandate given by the US Congress to the National Centre for Missing & Exploited Children (NCMEC) can be a best practice.
> The legal framework related to freedom of speech in the US and the protection offered to online platforms

[1] NP can be explained as the set of all decision problems (yes/no type) that can be solved in polynomial time by a non-deterministic Turing Machine and for which there exists a polynomial-time verification algorithm.

> enable the platforms to regulate illegal content online on their own, through the use of their Terms of Service/Terms of Use.
>
> The above statement is a result of the First Amendment not binding online platforms, and Section 230 of the Communications Decency Act (CDA) providing online platforms a broader immunity from liability for user-generated content posted on their websites.

Alas, every rose has its thorn. According to the same study, the *self-regulating* policies can create an environment with a lack of direct accountability to its users by decreasing transparency. Buni and Chemaly [2] argue that the organizational motives for the same are gaining cover from liability, safeguarding proprietary technology, preventing the gaming of their platform, and flexibility to respond to situations (likely due to a shortage to moderators).

In 2015, in a bid to improve transparency and make HN self-regulating, their moderators launched a fascinating experiment called Modnesty I. VentureBeat, had the following to say about this experiment [8]:

> Currently, when an account is banned, a software filter trips, or enough users flag a post, the post goes [dead], meaning only users with 'showdead' turned on in their profile can see it. At issue are posts that are incorrectly labeled as [dead] when they shouldn't.
>
> What Modnesty does is recall the [dead] posts on a case by case basis. Users who have at least 30 karma points will see a "vouch" link next to posts to revive them, but doing so will require support from the community. Just one vote won't do anything. All vouched posts will be reviewed by the administrators before reviving to verify that they don't violate HN guidelines.

The *showdead* feature of HN distinguishes a killed post from a permanently deleted one, and stems from a policy explained by Paul Graham in his 2009 essay titled *What I've learned from Hacker News* [10]:

> I think it's important that a site that kills submissions provide a way for users to see what got killed if they want to. That keeps editors honest, and just as importantly, makes users confident they'd know if the editors stopped being honest.

Based on what we have seen so far, it is reasonable to wander off and ask ourselves - What *approximations* apart from the constitution-mandated ones can these websites use to flag discussions? While *civility* certainly encapsulates all of them, on a more fundamental level, issues such as trolling, spamming, adult content, people/animal exploitation, hate speech, misinformation, harassment, and graphic violence are often looked out for. A curious reader can explore Pinterest's surprisingly transparent and non-archaic community guidelines [9] detailing the same. However, as argued by Martin Kleppmann [11], while dealing with *civility*, care has to be taken as *tone policing should not be a means of silencing legitimate complaints*.

Apart from these approximations, HN follows the underutilized *Broken Windows Theory*, which is described by Paul Graham in the same essay [10]:

> It's pretty clear now that the broken windows theory applies to community sites as well. The theory is that minor forms of bad behavior encourage worse ones: that a neighborhood with lots of graffiti and broken windows becomes one where robberies occur. I was living in New York when Giuliani introduced the reforms that made the broken windows theory famous, and the transformation was miraculous. And I was a Reddit user when the opposite happened there, and the transformation was equally dramatic.

The idea is that if we constrain our discourse model to dinner table values, we can promote more substantial and thoughtful discussions. However, does it take away the charm of a virtual setting? I leave this point for you to wonder about, hopefully at your dinner table.

Lastly, let us discuss a nonconformist opinion that requires greater community attention. Assuming communities to be legally valid, participation in one community should not automatically ban users from other communities. As painfully articulated by an HN user [12], it can be *detrimental to the discovery and growth of smaller communities that are viewed as controversial by bigger ones*.

III. The House of Cards: Moderation Architectures

> I would define online forums, newsgroups, and IRC/IM chat as predecessors to the current generation of social media, which are focused on the individual profile.
>
> On social media you follow people, while on forums you follow topics.
>
> - Anonymous on HN

On an elementary level, a moderation architecture consists of a group of webmasters overseeing a system with certain moderators arbitrating the discourse using special privileges. While the privileges may differ, most moderation systems provide anti-abuse software and user flags. Dissecting this anti-abuse software, some of the commonly found elements include shadow-banning, fact-checking, downweighting (like de-emphasizing keywords), and voting-ring detection [2]. Shadow-banning is primarily used for spammers and trolls, wherein the system bans the users without their knowledge. However, as rightly mentioned by Drew DeVault [13], care has to be taken that this feature does not become *the first line of defense against rulebreaking users*. Why? For

[2] A captivating albeit imperfect solution for this includes assigning a HyperLogLog counter [14] for every account, and updating it when another unique user upvotes content made by this parent account. We then find out the overall estimated count of unique upvotes and divide it by the total upvote counts for the parent user. Organic accounts are likely to have a ratio close to one.

any environment, to achieve transparency and self-regulation, it is a bare necessity that users are explicitly informed about their violations and are provided with a platform to appeal.

Slashdot, a social news website [15], consists of unique user-based architecture, wherein randomly selected users are periodically assigned certain points, which they can then use to rate the content. A comment is capped between -1 to 5 and users have the option to hide all the content less than a chosen threshold. Furthermore, in a bid to improve transparency, moderators have the choice of labelling content as *off-topic*, *troll*, *insightful*, *underrated*, etc. This can help understand the psychological or emotional state of a moderator's decision. Can such fundamentals help decentralize Facebook's *Oversight Board* model [16]? Certainly, an interesting argument to consider.

Continuing our discussion on decentralization, let us board a flight of imagination and ask ourselves - what would happen if there was a per-user feed manipulation? Say, the users had the choice of selecting a filter from a buffet of filters? Matrix, an encrypted decentralized open-source network, recently recommended a similar moderation system [17]. To prevent users from selecting a locally optimum filter bubble, they proposed a UI to visualize and warn users about the extent of their filtering. In such a setting, the admins running the servers can then decide the per-jurisdiction rules regulating their platform.

For many discourse models, meta-threads can act as a decisive regulating component by providing suggestions to mold and improve the system. Interestingly, HN has a different take on meta-discussion, wherein the site and its moderators actively discourage such content, correlating it with a *fast growing weed*. My understanding is that meta-discussions can at times lead to micromanagement, which may violate a key principle of HN (as explained in Paul Graham's essay):

> Every time the site gets slow, I fortify myself by recalling McIlroy and Bentley's famous quote "The key to performance is elegance, not battalions of special cases" and look for the bottleneck I can remove with least code.

Another key issue that plagues QnA and news-based models are duplicate content. As explained by a Quora engineer, having a canonical page can help users access all the answers in a single location and can connect writers to a broader viewership. In a bid to make this an open problem, Quora released a question-pair dataset in 2016 [18]. Moreover, a news site can moderate such content by crafting a weak duplicate link detector to provide sound content a new life.

While many of the above architectural tools have an aura of pessimism surrounding them, an optimistic moderation tool can be the *second chance pool*. Famously implemented by HN, this tool can help up-weight content which was originally unable to reach a wider viewership.

IV. Is there a winner?

> Commercial content moderation is not a cohesive system, but a wild range of evolving practices spun up as needed, subject to different laws in different countries, and often woefully inadequate for the task at hand.
>
> - Buni and Chemaly

While discussing this topic in a report titled *Who Moderates the Social Media Giants?*, Paul Barrett (Professor at NYU) provided a list of recommendations for social media giants such as Facebook and Twitter [19]. He postulated that the content moderators in these organizations need to be increased in strength, selected from a diverse list of countries (to improve awareness of local cultures), and be treated as full-time employees with access to high-quality wellness counseling and psychiatric treatments. Furthermore, he argued that expanding the fact-checking system by hiring experienced journalists can help curtail misinformation and showcase legitimacy. Interestingly, his report argues that a metric that identifies the *frequency with which deleterious material is viewed, even after moderators have tried to weed it out* can be used by government bodies for regulating social media organizations.

In conclusion, while I have tried my best to paint a nuanced perspective, as is shamefully evident at times, this essay is biased towards *the Hacker News model*. By striving for a simple UI with a lack of social-media elements, an almost self-regulating recommendation system, and a commitment to provide a platform for civil discussions, Hacker News managed to skillfully address some of the shortcomings I faced while using discourse models such as Quora [3] and Facebook. Alas, what works for one fails for another. Lastly, as is true for many things in our life, it appears that in a sea of uncertainty, change is the only constant.

References

[1] "Why HN is the way it is, and why we hope it will stay that way: Hacker news," Hacker News, 13-Oct-2021. [Online]. Available: https://news.ycombinator.com/item?id=28853335. [Accessed: 06-Nov-2021].

[2] Buni, C., &; Chemaly, S. (2016, April 13). The secret rules of the internet. The Verge. Retrieved November 6, 2021, from https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

[3] Graham, P., & Gackle, D. (n.d.). Welcome to Hacker News. Hacker news: Welcome. Retrieved November 6, 2021, from https://news.ycombinator.com/newswelcome.html.

[4] Altman, S. (2015, April 3). New Hacker News Guideline. Y Combinator. Retrieved November 6, 2021, from https://blog.ycombinator.com/new-hacker-news-guideline/.

[5] Decentralised content moderation: Hacker news. Hacker News. (2021, January 14). Retrieved November 6, 2021, from https://news.ycombinator.com/item?id=25776124.

[6] Aaronson, Scott. "Why philosophers should care about computational complexity." Computability: Turing, Gödel, Church, and Beyond 261 (2013): 327.

[7] DE STREEL, Alexandre. "Online Platforms' Moderation of Illegal Content Online." Retrieved November 6, 2021, from https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)652718

[8] Yeung, K. (2015, September 29). Y combinator spins out hacker news to give it 'full editorial independence'. VentureBeat. Retrieved November 6, 2021, from https://venturebeat.com/2015/09/29/y-combinator-spins-out-hacker-news-to-give-it-full-editorial-independence/.

[3] The biggest shortcoming while using Quora was the severely overfitting nature of their recommendation system and their lack of commitment to archive their answers. In my honest opinion, it would be an absolute crime if gems like Prof. Richard Mueller's answers are lost in the history of time.

[9] Pinterest. (n.d.). Community guidelines. Pinterest Policy. Retrieved November 6, 2021, from https://policy.pinterest.com/en/community-guidelines.

[10] Graham, P. (2009, February). What I've learned from Hacker News. Retrieved November 6, 2021, from http://www.paulgraham.com/hackernews.html.

[11] Kleppmann, M. (2021, January 13). Decentralised content moderation. Martin Kleppmann's blog. Retrieved November 6, 2021, from https://martin.kleppmann.com/2021/01/13/decentralised-content-moderation.html.

[12] Combating abuse without backdoors: Hacker News. Hacker News. (2020, October 20). Retrieved November 6, 2021, from https://news.ycombinator.com/item?id=24826951.

[13] DeVault, D. (2017, September 13). Analyzing HN moderation & censorship. Drew DeVault's blog. Retrieved November 6, 2021, from https://drewdevault.com/2017/09/13/Analyzing-HN.html.

[14] Berryman, J. (2013, October 14). Detecting reddit voting rings using this weird little data trick. Open-Source Connections. Retrieved November 6, 2021, from https://opensourceconnections.com/blog/2013/10/14/detecting-reddit-voting-rings-using-hyperloglog-counters/.

[15] Wikimedia Foundation. (2021, October 22). Slashdot. Wikipedia. Retrieved November 6, 2021, from https://en.wikipedia.org/wiki/Slashdot.

[16] Facebook. (n.d.). Governance. Oversight Board. Retrieved November 6, 2021, from https://oversightboard.com/governance/.

[17] Matrix.org. (2020, October 19). Combating abuse in matrix - without backdoors. Retrieved November 6, 2021, from https://matrix.org/blog/2020/10/19/combating-abuse-in-matrix-without-backdoors.

[18] Quora question-pairs data. Retrieved November 6, 2021, from https://www.kaggle.com/c/quora-question-pairs/data.

[19] Barrett, Paul M. ”Who moderates the social media giants.” Center for Business (2020).