# Ethical Questions for Volunteer-Reliant Moderation Systems

Joseph Seering
Stanford University
Computer Science Department and HAI
seeringj@stanford.edu

There have been a number of arguments made by scholars who study moderation recently, perhaps partly in response to the attention on /r/wallstreetbets, that suggest that relying on volunteers to do moderation is an inherently unethical thing for companies to do. Volunteer-reliant moderation may seem at first like a strange arrangement – many large companies spend huge amounts of money developing advanced algorithms and hiring armies of contractors to do moderation, while others seem to simply offload this responsibility on to users. On its face, this might reasonably seem like an unethical burden to place on users. In my research I've studied volunteer moderation on several different platforms, discussing these exact questions with more than one hundred volunteer moderators who moderate groups and communities on four different platforms,[1] and I thought I might offer some critiques of these arguments. The point of this essay is not to argue that all moderation systems that rely on volunteer labor are ethical, though I believe that certain approaches definitely can be, but rather to argue that the ethical questions are more nuanced than what has been reflected in the public debate. As part of writing this, I've also received feedback from a number of moderators on the points that I make here, and I denote (with *) the points that have been most influenced by this feedback.

The following are the main arguments I've read, along with some initial critiques of each:

**Argument 1: Companies that rely on volunteer moderators are offloading their responsibilities on to users.**

If we are to explore the above argument, there are a few considerations we need to make before jumping into debating its merits. First and foremost is the fact that there is a very wide variety of platform models that integrate volunteer moderator labor in different ways. The Wikipedia model is on one end, where users do all of the work and there is no centralized platform that profits from this labor in the same way that, e.g., Facebook profits from the labor of group moderators. On the other end (more or less) is (or was, depending on how it is re-implemented) the Parler model, where volunteers are expected to do pretty much everything. The vast majority of volunteer moderation systems fall somewhere in the middle, but the most common model is one where platforms and volunteer moderators have different but complementary responsibilities. This model is typically the one adopted by platforms based around groups (e.g., Reddit, Twitch, Discord, and Facebook groups). In this model, platforms are responsible for handling much of the moderation of the worst and most dangerous content (e.g., child sexual abuse, botnets, and state-sponsored influence campaigns) and also making the large-scale decisions like setting site-wide policies and deciding whether or not to ban entire communities. Volunteer moderators are responsible for handling the day-to-day moderation of the content posted to their communities, which requires active attention and awareness of group norms and the immediate social context of behaviors within the group.

According to the moderators I've spoken with and interviewed, this division of responsibility tends to work fairly well until the companies fail to make the big decisions they need to make or fail to support the moderators when they need support (technically or otherwise). One clear example of the former was Reddit's failures in dealing with /r/The_Donald, the main space for Donald Trump supporters on the site, a situation where Reddit CEO Steve Huffman even publicly admitted afterward that they had been far too slow and timid (my paraphrasing) in responding.[2] It's worth noting though, that the major volunteer-moderator-reliant platforms were actually faster and

more aggressive in dealing with problematic Trump-affiliated groups than their non-volunteer-reliant counterparts; Twitch suspended Donald Trump's account in mid-2020 and removed some of the recordings of content it had made from the platform. Reddit quarantined /r/The_Donald in mid-2019, a move that fairly quickly led to that subreddit's death. Though Discord's actions are less publicly visible, it has been fairly aggressive in removing far-right spaces since at least mid-late 2017. So – typical volunteer-moderation-reliant systems aren't simply about offloading responsibility; they're best understood as a system that creates a division of responsibility according to which party is best equipped to handle which type of task.[3]

I think the biggest counterpoint to Argument 1, however, is that the volunteer moderators actually generally want to have the responsibilities that they have. I don't think it's a particularly extreme assertion to say that if, e.g., Reddit were to announce that they had, e.g., "realized that we have been exploiting volunteer moderator labor, so we will be taking responsibility for all moderation-related tasks from now on," users would leave the platform en-masse and it would probably collapse. One interesting and fairly direct example of this was when Reddit tried to introduce a centrally-moderated chat feature in mid-2020 that would appear on subreddits but would be moderated by the company rather than subreddit moderators. A significant number of moderators rioted (more or less), and the feature was withdrawn within 24 hours and an apology from the company was put forth.[4] One of the main attractions of Reddit is that communities are run by people who typically care a lot about the subjects that are the focus of each subreddit. The moderators of /r/buildapc know how to build custom PCs. The moderators of /r/Esperanto know Esperanto. The moderators of /r/ITCareerQuestions/ are actually IT professionals. These moderators can make moderation decisions based on deep contextual knowledge of situations that come up. It doesn't always work out that way – humans are imperfect – but these moderators are far more in tune with their local contexts than Reddit-hired moderators ever could be. An interesting thing I've seen recently is that Facebook groups will often have a moderator-written rule that directly states some variant of "Do not use the report button to report content in this group to Facebook. Tag a moderator and we will take care of it. Facebook will just screw things up." These are not extremist groups – the ones I've observed are mostly just leftist groups where arguments sometimes get out of hand before volunteer moderators have a chance to step in.

I think we have to see volunteer moderation as more complicated than "companies offloading their responsibilities on to users" when users will go out of their way to keep these responsibilities for themselves. The ethics of volunteer-reliant models for moderation must be considered in context of the broader set of models. *Is it really more exploitative to expect users to manage the social spaces that they care about than to prohibit them from doing so?* As a final note, if we're critiquing volunteer-reliant moderation systems, it's really important to critique real, implemented systems and to consider their actual features and processes as part of our analysis. If we do so, we can see that, e.g., there are many critiques that apply to Parler's model that don't apply to Reddit's (and perhaps vice-versa).

**Argument 2: Relying on volunteers to do content moderation is unethical because it exploits user labor.**

I think this is probably the most interesting argument that I engage with, and it's one that I hear often from people who are familiar with the basics of volunteer moderation systems. I think this issue overlaps at its core with the broader question of platforms profiting from user-generated content; social media platforms simply could not exist if users didn't generate content that other users are interested in seeing. For better or for worse, I've seen less discussion of this UGC issue recently. It seems like the public (and scholars) have mostly accepted that this is a fair model, but reasonable arguments have been made in the past that platforms that profit off of UGC are ethically obligated to reinvest those profits in a way that benefits content creators, either directly by paying them or indirectly by building features that help them create better content. I don't necessarily agree with this argument, but I think it may be a valid position to take. There's a parallel argument to be made about volunteer moderation that I do personally support – that companies that rely on volunteer moderation are ethically obligated to be responsive to the concerns of these moderators both in terms of communicating with them and collaborating with them in some form to develop tools to better facilitate their labor. I think this latter argument matches well with the above response to Argument 1, and also resonates with what I've heard when talking with volunteer mods across multiple platforms. Their biggest complaints are definitely not that they're expected to do the labor of moderation, but rather that the company is not sufficiently responsive to their concerns and/or that the tools provided are inadequate for

handling certain specific situations.

**Argument 3: Relying on users to do content moderation is unethical because it forces users to be exposed to potentially-traumatic content.**

There is certainly some truth in the assertion that some volunteer moderators on major platforms are sometimes exposed to potentially-traumatic content. Moderators who I've interviewed and spoken with have occasionally described such incidences. The moderators who this happens to are, more often than not, in the small subset of moderators who moderate very high-volume spaces. The first note I will make in response to this argument is that, per my response to Argument 1, volunteer moderators on major platforms are not, in practice, simply thrown to the wolves to deal with the worst of the worst. Platforms like Reddit and Discord use algorithms that function similarly to the ones used by Facebook and Twitter to actively remove at least some categories of extreme content; to the best of my understanding, the teams at these companies tend to feel that the responsibility for moderating these types of content should fall on the company rather than the volunteer moderators, and these teams do make an effort to adhere to this responsibility (albeit imperfectly). Moderators I've spoken to also say that, when they encounter a very extreme behavior or type of content that they are not comfortable dealing with, they will report it and request that the platform deal with it, and these platforms are often, though not always, fairly responsive to these requests. I think the better argument* to be making here (as I note below) is that volunteer moderators may not be adequately prepared for the responsibilities they're signing up for, and may not have adequate resources to deal with potential trauma when it does occur. I've written in a couple of papers that I don't think companies do enough to onboard moderators, and I think this continues to be true. The Discord Moderator Academy[5] is one way in which a major platform is trying to improve on this front, but I think more can be done on both Discord and other major platforms.

**Argument 4: Volunteer moderators are less accurate than professional moderators.**

This argument puzzles me because it is often made by the same scholars who are (in other contexts) sensitive to questions about the situational nature of "accuracy" in human judgment.[6] Questions about moderation are, in nearly all cases, subjective decisions. How, practically-speaking, are we supposed to evaluate the "accuracy" of volunteer moderators' determinations of what specific pieces of art are too sexual to be permitted within a group dedicated to creative expression? Many essays and papers have been written about the deep nuances of similar situations companies have encountered. In cases like these, where moderators make judgments about what behaviors cross subjective lines, volunteer moderators who are sensitive to local contexts are arguably likely to be *more* "accurate", not less. Somewhat more broadly,* I think it's fair to argue that no social media company could ever hire enough professional moderators to become familiar with the social context of every community. The number of moderators required would be more than even Facebook can afford.

From my research, I've found that decisions like these are a substantial fraction of the decisions volunteer moderators spend their time thinking about. There are certainly "routine maintenance" tasks that are more black and white in terms of "accuracy" (e.g., removing scam links), but the important decisions tend to be the ones that fall into the former, more subjective category.

**Argument 5: If anyone can be a moderator, communities can develop where people are allowed to behave in terrible ways.**

A fairly common argument against reliance on volunteer moderation is that it is inherently morally relativistic; volunteer moderators can have any value sets, and thus they may be okay allowing lots of content to be posted that violates broader social norms. For the tiny subset of platforms that rely exclusively on volunteer moderator labor, this can certainly be true, but most major platforms operate under the two-tiered, split responsibility model that I discussed above. When volunteer moderators allow (or even facilitate) problematic communities, the platform can step in to remove those communities. This happens fairly regularly on all of the major platforms that rely on volunteer moderator labor. There are fair arguments to be made (that I generally agree with) that these platforms (e.g., Facebook and Reddit) are not proactive enough in removing problematic communities, and we have seen, e.g., Reddit grow (agonizingly slowly) more aware of its responsibility in this regard over the last decade, but this

is a critique of the company's execution of the model, not an inherent flaw of the model itself.

—

With all of that said, here are what I think are some thoughtful questions that could be asked that could help us improve the space of models for volunteer-reliant moderation systems:

**1. Who should be allowed to be a moderator?* E.g., should minors be able to be volunteer moderators? How might this vary across different spaces?**

In the current system on most of these platforms, the moderator of the group you're in could be a fourteen-year-old, a convicted murderer, or a dog, and you might not have any way to know. Should there be some check on these systems to limit who is allowed to be a moderator? Should these vary across spaces? Should we, e.g., check to make sure no volunteer moderators of groups for teenagers are convicted sex offenders?

**2. How can volunteer moderators be adequately supported by platforms?***

Typically I group the ways companies can support moderators into one of two categories – communication and collaboration. In the first category, I believe that platforms have some obligation to listen to moderators to get a better understanding of their values and their needs (both in the short term and the long term), and should show that they have listened and understood. I also believe that platforms should collaborate with users to develop the tools and interfaces that will allow them to better handle the challenges they face in moderation. This can probably be done through established participatory design and user-centered-design methods.

**3. How can volunteer moderators be appropriately compensated for their labor?**

This question relates to Argument 2 above. A first response to this question would be that, because platforms profit from volunteer labor, these volunteers should be paid for their work. On a platform like Facebook, which has literally millions of users volunteering their time to moderate groups, this is not likely to ever happen. However, there are other ways of compensating users, ranging from simply showing more respect for them and the value of what they do to explicitly granting them extra privileges based on what they do. One moderator I interviewed said that it would be nice if, e.g., Reddit granted "Reddit gold" to moderators as a small token of appreciation. This is typically the level of "compensation" that moderators I've spoken to seem to want – acknowledgement, respect, and support.

—

I have a lot more thoughts on all of the above, but I've endeavored to make my summary of the major points as brief as possible. I'd be happy to engage with anyone who wants to take the time to write up a response, and I hope the above has been interesting enough to prompt people to think more about these questions.

*These points were prompted by conversations with volunteer moderators

## NOTES

[1]The works that I've authored or co-authored that most closely relate to these points are "Moderator Engagement and Community Development in the Age of Algorithms" (2019) and "Metaphors in Moderation" (2020), published in *New Media & Society*, and "Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation" (2020), published in the Proceedings of the ACM on Human-Computer Interaction Vol 4, CSCW.

[2]https://old.reddit.com/r/announcements/comments/gxas21/upcoming_changes_to_our_content_policy_our_board/ft07637/

[3]Economists might relate this to the concept of *comparative advantage*; I think this is an imperfect but interesting comparison.

[4]https://old.reddit.com/r/ModSupport/comments/gafm52/mods_must_have_the_ability_to_opt_out_of_start/fp0r557/

[5]https://discord.com/moderation

[6]Some Computer Scientists have even started to discuss the problems with treating decisions made by moderation algorithms as objectively "correct" or "incorrect"; a fascinating recent paper by a colleague of mine at Stanford discusses this at length from a mathematical perspective: https://hci.stanford.edu/publications/2021/gordon_disagreement/chi2021_disagreement_deconvolution.pdf