

Reporting System Rubrics: Summary

Summary of research

Also available as a detailed report.

Claudia Lo, Design Research, Anti-Harassment Tools Team
For the Community Health Initiative, Feb 2019

The content contained in this publication is available under the Creative Commons Attribution-ShareAlike License v3.0 (<http://creativecommons.org/licenses/by-sa/3.0/>) unless otherwise stated. The Wikimedia logos and wordmarks are registered trademarks of the Wikimedia Foundation.

Use of these marks is subject to the Wikimedia trademark policy and may require permission (https://wikimediafoundation.org/wiki/Trademark_policy).

Introduction	2
Designing a rubric	2
Reddit	3
Users	3
Moderators	4
Conclusion	5
Facebook Groups	6
Users	6
Moderators	7
Conclusion	8
Takeaways	9

Introduction

On Wikipedia, most content and conduct disputes are handled by groups of volunteers. Accordingly, reports of such disputes are first routed to them, and only in cases of immediate danger or outsized harm do reports bypass this volunteer system and go directly to the Foundation’s Trust & Safety team. At this stage in our ongoing project on creating private reporting systems for Wikipedia and other Wikimedia projects, we could learn from investigating peer platforms’ implementation of private reporting systems. As our editors access multiple online platforms in their digital lives, other platforms’ reporting systems will inform their expectations, and so should be considered for our own future designs.

By conducting a review of existing best practices documents and research on this subject, we can create an assessment rubric to evaluate private peer-to-volunteer reporting systems. Some of the most prominent platforms using such a system include Reddit and Facebook Groups. We can run these platforms through this rubric, and additionally compare the current state of Wikipedia’s reporting systems, for a comparative understanding of these mechanisms.

Due to the length of this investigation, this report has been split up into a presentation slide deck, a summary report, and a detailed report.

Designing a rubric

The rubric looks at four major areas: accessibility, ease of use, communications, and privacy. There are two versions of the rubric for each platform, one from a user’s perspective, and one from a moderator’s perspective.

Each quality being assessed can be rated as “complete”, “partial”, or “sparse”. Complete does not necessarily mean good, nor does sparse mean bad. These categories are meant to illuminate the design priorities of these systems; for some categories, one could hypothetically find fault with both a “complete” implementation of the quality, or the “sparse” version.

Finally, this rubric was designed to assess only the technical reporting system, the mechanism by which a user could make a report and send that report to a volunteer moderator. It is not meant to take into account social practices: for example, though it is common practice on

English Wikipedia to notify a user mentioned in a report, it cannot be done through the actual process of reporting, which is simply writing a report in an open text field.

Reddit

One important detail to keep in mind is that Reddit is currently undergoing a thorough redesign. This means that some pages, mostly the ones for users, have been redesigned and now follow a generally coherent visual style. Additionally, this overview does not cover Reddit's new chat feature, since reports on that system goes directly to staff instead of volunteers.

Users

Reddit user rubric	Complete	Partial	Sparse
Accessibility	1	4	1
Ease of use	1	2	7
Communication	0	2	3
Privacy	2	1	4

Users can report individual comments or posts, using a “report” link found on every post or comment, which is also present on mobile. There is no way to report a specific user. These reports go to subreddit moderators, but nowhere is this indicated on the reporting form.

Official documentation on reporting is sparse and users are not explicitly introduced to it in onboarding, leaving volunteer moderators to pick up the work of informing their users how to report and what should be reported.

The reporting form is quite easy to use, with many pre-filled options as well as a 100-character custom response option. Only one option can be chosen at a time. A user can only report once per post or comment; while they can continue to go through the form, subsequent reports go

nowhere. A user cannot edit a report, once filed. The mobile experience is functionally identical to the desktop reporting system.

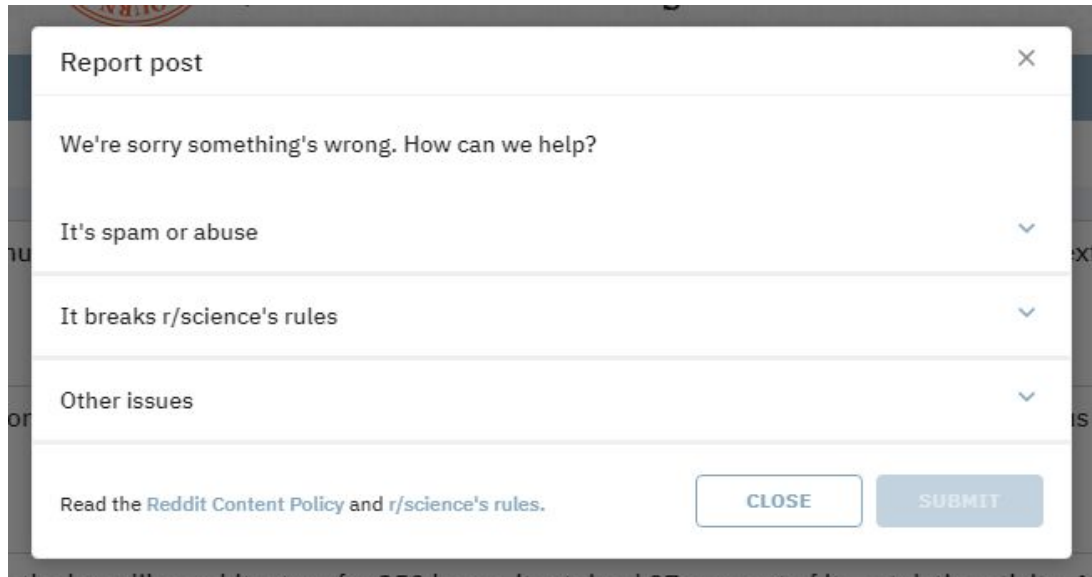


Fig. 1. The reporting form, for r/science.

Due to the system's design, users cannot be contacted directly about reports that they have filed. Reddit's private messaging system allows a user to send a private message to all the moderators of a subreddit, using the modmail system. However, whether or not a user is informed about the status of their report relies on them voluntarily telling moderators that they have made a report on a particular post, and for the moderators to willingly tell the reporter about any actions they take. Moderators cannot initiate this conversation.

All reports on Reddit are anonymous by design, though users must be logged in to report. Users can report on someone else's behalf, but cannot specify the person for whom they are reporting. Reports are never associated with their reporter and are always kept private, but none of this is indicated in the form.

Moderators

Reddit mod rubric	Complete	Partial	Sparse
Accessibility	2	3	1
Ease of use	8	5	3
Communication	0	5	2
Privacy	2	2	2

Any user could become a volunteer moderator on Reddit. The relevant permissions for moderators also grant access to subreddit traffic reports and the mod log, which records all actions taken by the moderators of a subreddit.

Moderators can access all reports via a central dashboard, called the modqueue. All reports are grouped with the reported post or comment, and are sorted chronologically, newest at the top. Reports cannot be sorted any other way. They are generally clear and easy to read, but only display short text. On mobile, moderators have access to the same sets of actions as they do on desktop.

Reddit allows for some automation via AutoMod, which can generate reports but cannot be used to address them. There are third-party tools developed to help moderators handle reports. Relevant information is clearly presented, but some potentially useful information, such as timestamps on reports, do not exist. Conflicts are handled in a “last action wins” way, where the last action to be taken overrides all others. Reddit’s guidelines for moderators deal with general best practices, but not with the specificities of handling reports or sensitive issues such as harassment.

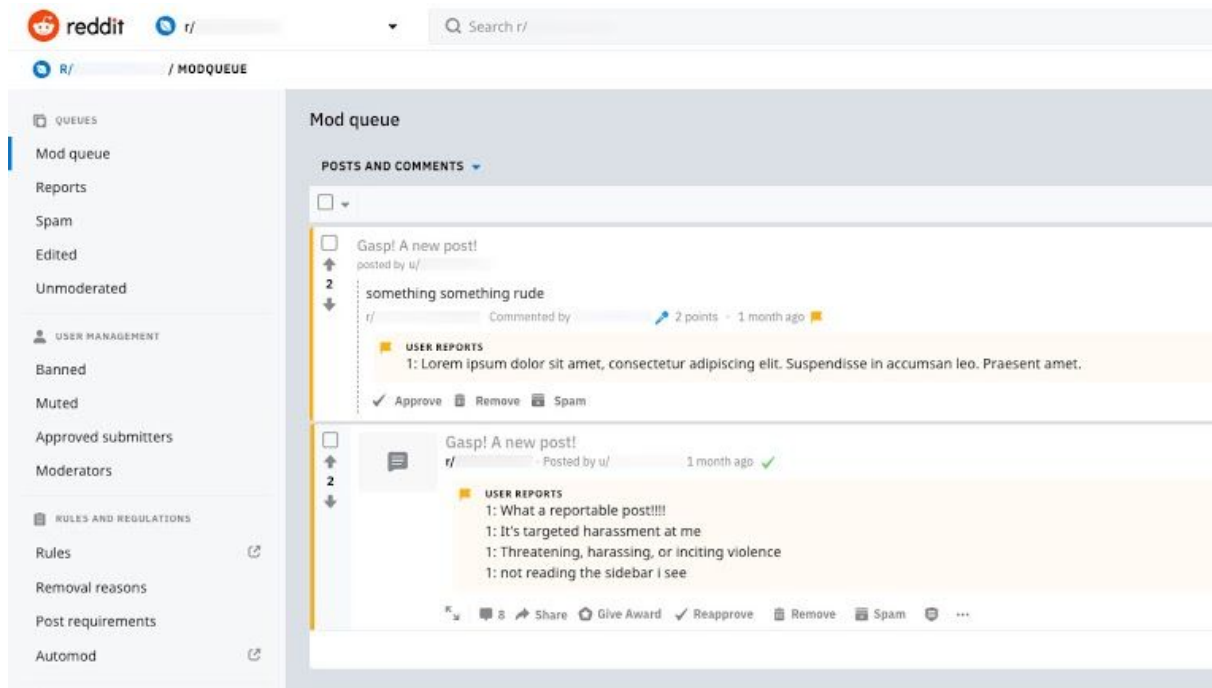


Fig 2. A screenshot showing the mod queue; the option to ignore reports has been truncated, where it would normally be on the right of each reported post or comment.

Escalating reports to other groups is very difficult, since reports are attached to a specific post, and there are no specific pathways for moderators to escalate a heavily-reported post. They must make use of the same methods as every other user, namely emailing Reddit or privately messaging r/reddit.com with the built-in messaging system. Communications in general can be challenging, since moderators cannot tell who has made a given report.

Once a report has been acted upon, either by ignoring it or removing the reported content, it can be tricky to reopen reports. Ignored reports can still be accessed, but a removed-then-reapproved-then-reported post or comment will not show up in modqueue, rendering it practically invisible. Moderators can leave persistent notes for each other via a notes system, though these notes are also limited to 100 characters, and can be difficult to archive.

Moderator actions are logged in the modlog, visible only to other moderators. This is not publicly viewable. Lastly, all abuses of the reporting system are meant to go to Reddit staff, and the “ignore reports” function serves as one of the few anti-abuse measures this system has.

Conclusion

One of the strengths of Reddit’s reporting system is the immediacy of a reporting option, and the relative ease with which one can make a report. Additionally, while making all reports anonymous causes some troubles with regards to documentation, it does ensure the privacy of all reporters. While the moderator queue is undoubtedly useful, as is the ability to leave removal reasons and the ability to see rough chronological actions in the mod log, intra-mod communication on Reddit itself is still somewhat basic.

Ultimately, Reddit’s reporting system is very well suited to one-off instances of unacceptable content, and ill-suited for reporting either harassment in private messages or long histories of unacceptable behaviour.

Facebook Groups

Facebook, broadly speaking, relies on the use of commercial content moderators, often based far from the cultural contexts they are expected to moderate. Though Facebook is popularly thought of as not using volunteer moderators, Facebook Groups remains an exception. Unlike most of Facebook, Facebook Groups allows volunteers to act as moderators or administrators for the group, and these volunteers can choose to remove content or users from the group.

This report is only concerned with the reporting system for reporting within Facebook Groups. While this will overlap somewhat with Facebook’s site-wide reporting systems, it will focus heavily on the mechanism for reporting to group administrators and moderators.

Users

FB Groups user rubric	Complete	Partial	Sparse
Accessibility	1	4	1
Ease of use	1	3	6
Communication	0	2	3
Privacy	2	1	4

Reporting to Group administrators is different for posts and comments, though as with Reddit, there is no way to report specific users. It is quite simple for posts, since reporting is an option in a post’s breadcrumb menu. However, reporting a comment is a five-click process involving counterintuitive options. An administrator’s post in a group cannot be reported back to other administrators, only to Facebook. Administrators cannot report comments or posts, and can only remove them in response to objectionable content.

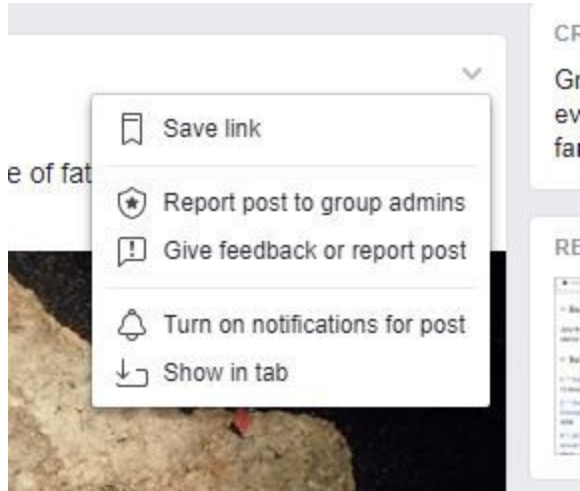


Fig. 3. A screenshot showing the reporting options available on a post on Facebook Groups.

Documentation on how to report posts or comments exists, but can only be found by searching. Otherwise, its quality and relevance relies on whatever documentation group administrators have produced.

The reporting form now allows users to specify a reason for reporting, but does not allow users to set custom responses or reference a group's own rules as a reason. The labelling of the report option makes it clear that a report to group administrators goes to them, not to Facebook. Reports only allow one reporting reason.

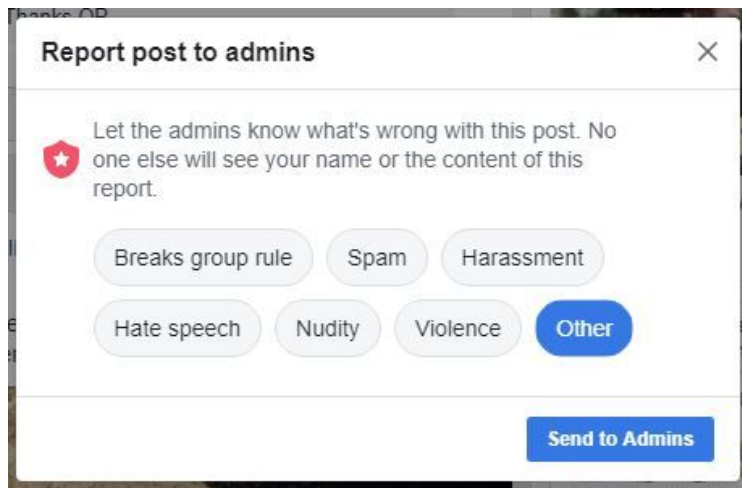


Fig. 4. The new (as of 9th Jan 2019) form for reporting a post to group admins on Facebook Groups, with the Other option selected.

Communications around reports are difficult. Once a report is made, it leaves a user’s control, especially for reported comments; because reporting comments requires the user first hide that comment, on a page refresh, the reported-and-hidden comment will disappear from view with no way to restore it. Users could message group administrators via Facebook Messenger, but since Messenger deprioritizes messages between users who are not friends, there is no guarantee administrators will see those messages in a timely manner.

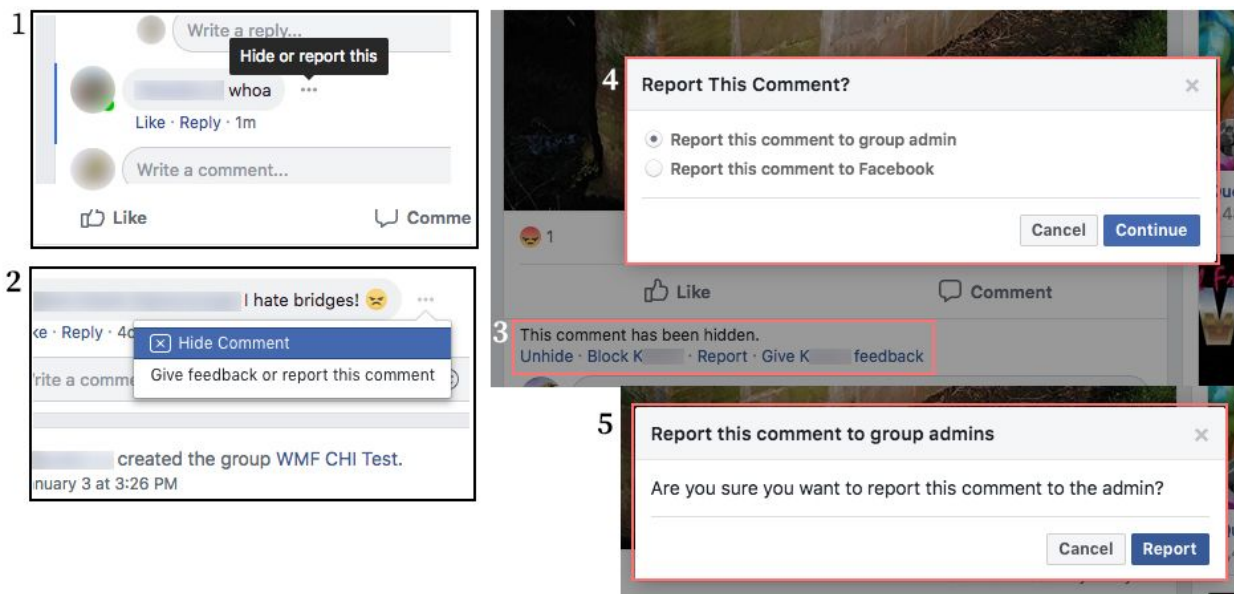


Fig. 5. The process of reporting a comment, on desktop. In numbered order: hovering over the breadcrumb menu (1), hiding the comment (2), the results of clicking “Report” (3) in the new line of links (highlighted with a red outline added by author) with a form specifying type of report (4), and a confirmation window (5).

Every report has the reporter’s username attached. Because of Facebook’s policies, this is likely to be personally-identifying information. All users must be logged in to report, and it is impossible to report anonymously or on behalf of another. Reports are kept private, but the reporting form does not tell users what kinds of data it captures along with the report. This means it is unlikely that a user, who is not already a group administrator, would realize their name is attached to reports that they make.

Moderators

FB Groups mod rubric	Complete	Partial	Sparse
Accessibility	2	3	1
Ease of use	8	5	3
Communication	0	5	2
Privacy	2	2	2

Users in charge of governing a group on Facebook are split between administrators and moderators. While they largely have similar permissions with regards to hiding or deleting content and controlling users' access to the group, administrators can additionally grant moderator or administrator privileges. Because different products on Facebook are handled differently, the effects of identically-named moderator actions can vary between products.

Moderators have access to a central dashboard and report queue, on both desktop and mobile. Moderators are notified when reports come in, and will see a banner alerting them to new reports in the queue. All reports are attached to their post, sorted in chronological order according to the age of the reported posts. Reports cannot be further sorted.

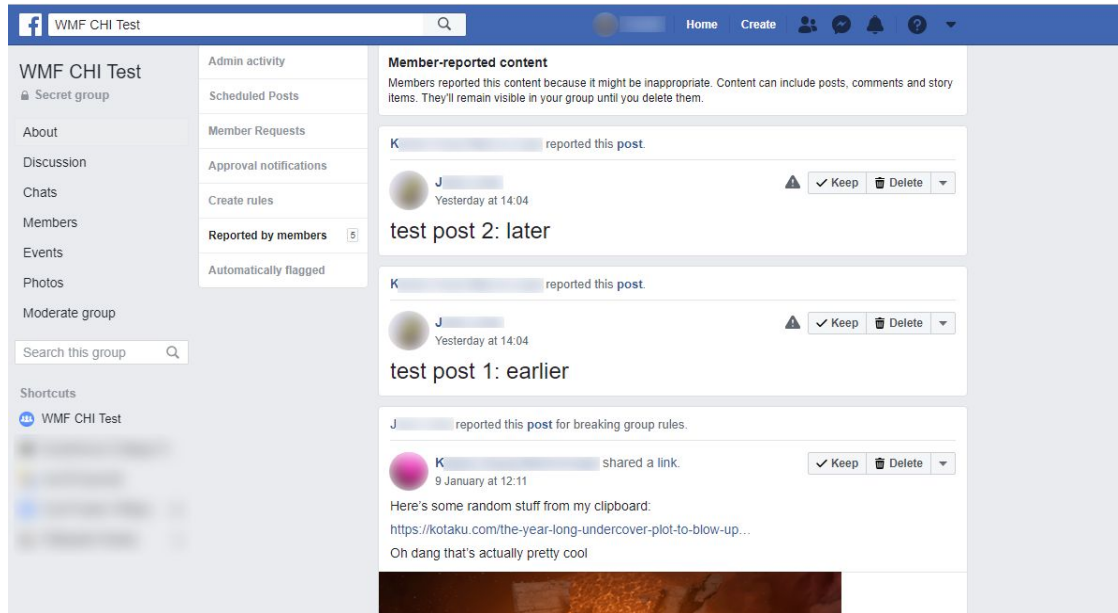


Fig. 6. A screenshot of the reported items queue.

Facebook Groups does not support automated tools or third-party extensions, nor does it have mass approval or removal tools. Though user-provided reasons are easy to read, they are subtle and easy to miss. The system provides the reporter's name, as well a count of how many moderator actions have been taken against the author of the reported content. Report reasons cannot be customized or reference the group's own rules. Conflicts are handled on a "last action wins" system, similar to Reddit.

It is very easy to access Facebook's guides to moderation, though their rapid and unannounced development means that these guides focus on general moderation guidelines and not the specifics of addressing reports. Removing a user from a group is a fairly granular process, but removing content is far blunter by comparison. Reinstating content is very difficult. There are no special escalation paths for moderators, but reporting directly to Facebook is extremely easy and so moderators can use that pathway to escalate relevant content to the company.

Most moderator actions are tracked in a log accessible only to other moderators. Report histories are not saved, and communications requires the use of Messenger, which has the pitfall of deprioritizing messages sent between non-friends. To get around this would require

friending involved users, which may give the unwanted appearance of intimacy or accessibility. Moderators can leave notes for each other on moderator actions that they take.

Due to the fact that usernames are captured along with reports and in the moderator log, moderators have access to personally identifying information of both reporters and moderators. Though no reports are publicly logged, there is another concerning factor not disclosed by Facebook: whether or not reports to group administrators are logged *by Facebook*. Moderators are not encouraged to employ additional security measures, and there is little anti-spam function built into the system. Additionally, the relatively flat hierarchy means that a single compromised administrator account could lead to significant problems for a group.

Conclusion

Facebook Groups' reporting system has a few notable strengths. For moderators, the system is fairly flexible in the breadth of possible actions for sanctioning users, and the system captures some useful information, such as number of moderator actions taken against a reported post's author. At the same time, its constant development shows that there is some level of investment put into developing the system.

However, its main weakness is the lack of clear communication when it comes to data visibility and functionality. As a consequence of such opacity, exacerbated by a rapid pace of unannounced development and constant A/B testing, users and moderators alike lose trust in the system.

Ultimately, this reporting system is very well suited for reporting a specific type of incident, that is, a flag on a post containing content that is clearly in violation of group rules, Facebook terms of service, or both. It is ill-suited for cases more complex than this, or for longer-term issues.

Takeaways

While the specific needs of a reporting system built for Wikipedia will of course be different to Reddit and Facebook Groups, there are still takeaways to be had from this assessment. One very important thing to keep in mind is that these platforms are teaching their users what to

expect of reporting systems on other platforms; one can reasonably assume a “report” option on Facebook, much like on Reddit or Twitter, will flag that post to some other group for review. There is no reason to assume this will not hold true for new editors’ expectations on Wikimedia projects.

Both Facebook Groups and Reddit rely on putting the “report” link in as many places as possible to make it visible. These reports are also all standardized. The major benefit is that, for common case, a standardized form greatly speeds up and structures the report. The drawback is that, for more complex cases or for cases that require context to explain, the lack of flexible reporting options like attaching media or free-answer text fields severely constrains the reporter’s ability to make a useful report.

Communications are not always thought of as part of a reporting system, yet escalation and mediation rely on easy and clear communication between involved users. Nor is it always clear where reports end up, or what happens to a report once it is made. Both of these mean that it is difficult to tell if, as a reporter, you are making any impact at all. At the extreme end, opaque communications can lead to distrust of the system, as we see in Facebook Groups. Given that we are designing a reporting system meant to handle potentially sensitive disputes, a lack of trust would severely hamper its effectiveness.

Lastly, we see a constant tension between balancing the desire for more information with the need to respect user privacy. Total anonymity and untraceable reports mean that the reporter’s privacy is always guaranteed, but makes it more difficult for moderators to resolve issues. However, attaching personally-identifying information to every report also seems unnecessary. The question becomes, how do we adhere to transparency in a way that is safe—both for reporters and the moderators handling reports—and respects the privacy of reporters?