Responsibilities of Privileged Users

Dr Andre Oboler and Adv David Matas
Online Antisemitism Working Group Co-Chairs
Global Forum to Combat Antisemitism

Background

The paper is an elaboration on recommendation D 20 of the 2009 Report of the Online Antisemitism Working Group of the Global Forum to Combat Antisemitism.[1] This paper has been prepared in response to a request for elaboration by Facebook's Richard Allan at the 2011 Working Group meeting.

Need

Online hate is a large and growing problem in social media platforms. The link between manifestations of hate on the Internet and real world hate crimes has been highlighted by the OSCE's Office for Democratic Institutions and Human Rights (ODIHR).[2] New approaches are needed in light of Web 2.0 technologies, and social media in particular.[3] Technical difficulties have also been identified due to the volume of data.[4]

Solution

With power comes responsibility

In social media platforms users with a privileged positions and the ability to remove other people's content, or to exclude another user from a space, should have a reciprocal responsibility to use their authority to remove problematic content which would otherwise need to be removed by the platform itself.

By sharing responsibility with the user community, the platform providers lower the burden their staff currently face in directly responding to this problematic content. The authority needed is no higher than currently exists. Repercussions beyond removing the content or excluding the user from that particular page / group remain at the discretion of the platform provider.

Beyond cleaning up the platform, policies such as this promote good citizenship and educate users to stand up against hate speech.

Empowering platform providers

Exception reporting showing users whose actions multiple administrators have seen fit to responded to may trigger a review by platform staff. Such user may face higher sanctions such as having privileges removed, or their account temporarily or permanently disabled.

---

[1] http://www.gfantisemitism.org/Conference2009/Pages/default.aspx
[2] Report of the OSCE-ODIHR Expert Meeting, Incitement to Hatred vs. Freedom of Expression: Challenges of combating hate crimes motivated by hate on the Internet, Warsaw, 22 March 2010 <http://www.osce.org/odihr/68750>
[3] Ibid
[4] http://www.osce.org/odihr/68743

In taking such action, the platform is not acting along, but is acting in response to the wishes of the user community who originally filed and judged the complaints.

Ensuring moderation occurs

For this system to work, privileged users must be able to anonymously see complaints lodged about content in spaces they control. It is also necessary for them to respond to each complaint with one of the following actions:

- Dismissing the complaint (no action)
- Accept complaint as a minor violation (and remove offending content)
- Accept complaint as a serious violation (and remove the user)
- Marking the report itself as abuse (report the person who sent the report)

If no privileged user responds within a reasonable time, a reminder should be sent to all privileged users for that space. The reminder should tell them they have pending complaints to review. If still no action is taken a warning should be issued telling them they risk losing their privileged status in the given space if no action is taken within a specified number of days. Following a second warning their access should be reduced to that of a regular user (with the exception of being able to review complaints) and they should be informed of this and warned that a failure to either review complaints or remove themselves as a privileged user in that space may result in system wide restrictions. Following a further period without review activity their privileged status should be removed within that space and their privileged access in other spaces suspended for a set number of days.

Each time the above process runs to completion, the suspension of privileges at a system wide level should be for a longer period. Another option would be to prevent repeat offenders from becoming privileged users in the future, however this removes the incentive for good behaviour even by repeat non-performers.

Penalties for the space can also be considered, such as disabling posting while there are no active moderators.

Ensure moderation is accurate

It is important to ensure moderation is not only occurring but is reasonably accurate. This can be achieved by a combination of random checks by platform staff and reviewing though exception reporting.

The reviewing through exception reporting would involve platform staff selectively auditing the actions of privileged users whose reporting is too far away from the average. This may indicate either an abuse of the privileged status or a privileged user acting responsibly in a space that is attracting a disproportionate level of problematic content. Either way this is a situation the platform provider should review.

The random reviews will catch other situations, such as spaces designed to spread hate where users are reporting those who complain and moderators are verifying those reports.

Random reviews can also be triggered when a privileged user is reported in their capacity as a regular user of the platform.

Benefits

This system will cause cultural change by educating users on the responsibility that comes with power, and on the need to uphold the platforms terms of service. It will also enable a far larger amount of problematic content to be removed far more quickly.

Evaluation

The change can be evaluated on a quantitative level by the number of items removed and also by the average turnaround time on complaints.

Costs

The above changes involve some development of the platforms. The staff needed to make the system work can be reassigned from their current work directly reviewing content, so there is no added staff cost.