Guardian News & Media (GNM)—supplementary written evidence (FEO0111)

How we moderate comments on our site.

This note is in response to a request from the Committee for detail on how the Guardian moderates comments on our services.

You may be aware that in 2016, The Guardian published a research project, called “The Web We Want”,[1] which sought to shine a light on the rising global phenomenon of online harassment. The project commissioned research into the 70m comments left on our website since 2006 and found that, of the 10 most abused writers, eight were women, and the two men were black. Of the 70 million comments left on the Guardian since 2006, and the commissioning of the Web We Want report in 2017, 1.4 million of those comments (2% of the total received) have been blocked by GNM moderators because they violated GNM’s community standards. Most of these are abusive to some degree (they may use insulting language or be ad hominem attacks) or are so off-topic that they derail the conversation.

At the heart of our approach to moderating comment sections on The Guardian, is an internal team of experienced moderators, dedicated to improving the experience for readers and staff visiting the comments section below the line on our website. Partly in response to the ‘Web We Want’ project, and as a way to better manage that dedicated internal resource, we sought to invest in technology that makes it easier for our dedicated moderation team to do their job effectively.

Recognising the role technology in reducing the workload of the moderation team, in 2016, a cross-functional team of colleagues from across the Guardian developed a machine-learning algorithm, affectionately known as Robot Eirene named after the Greek goddess of Peace, that supports the role of human moderators by spotting comments which contain personal abuse, author abuse and dismissive trolling which otherwise might be missed in unwatched threads.

The role of Robot Eirene does not replace human moderators, but rather it serves to reduce the volume of comments in our queues and to have high risk comments flagged to the moderation team. With a slightly reduced number of comments in the queues, human moderators are able to watch and pre-moderate more threads without significantly increasing their workload.

Two internal developers used machine learning techniques and our existing body of moderated comments to build a model that classifies comments as good or bad (where 'bad' is likely to require moderation). Robot Eirene is now used internally in two ways:

1. To identify 'bad' comments that require moderation but might otherwise not be seen by the moderators (22% of comments flagged are

[1] https://www.theguardian.com/technology/series/the-web-we-want

subsequently blocked - a much better rate than for abuse reports from users).

2. To identify 'good' comments that can be passed without moderation (0.3% of these are subsequently reported/blocked, suggesting a high level of accuracy here).

In addition to the development of this tool, we have evolved our approach to managing comment threads in the following ways:

- We have rationalised the number of threads we have open on the site with the explicit goal of enabling us to manage threads more effectively in order to maintain a good quality of discussion.

- We pre-moderate some threads to try and ensure that we can host a good discussion on more sensitive issues but we mainly post-moderate our comments.

- We also encourage staff to highlight good comments and engage with readers below the line, and seek to highlight good quality comments at the top of the thread.

- We send around a weekly comments update to editorial colleagues, highlighting the top staff commenters of the week, the best threads and comments generated throughout the week.

- In terms of the process to become a commenter on our site, all commenters have to register on the site. Users are allowed to use pseudonyms to protect their identity, to enable commenting on sensitive issues.

- When users sign up, they're signing up to strong community guidelines, and we're clear that where they make comments that breach those guidelines, they are subject to our moderation policy.

- To enable the team to spot trends in commenting behaviour, colleagues hold regular weekly team meetings where we discuss trends and tropes in comments and across media in general so we can keep ahead of what terms commenters on the site might be using to evade moderation.

- There is a review process for sanctioned readers to find out why their comments have been moderated which improves transparency and trust between the reader and the Guardian.

- Readers are able to find out why they have been sanctioned by contacting moderation managers. Users are able to appeal against sanctions by contacting the independent readers' editor.

In terms of enforcement of our policies, where a user makes a series of comments which break our standards, a sanction is put in place meaning that we then see their comments before they're posted on the site. We seek to enforce a one user, one account policy, and we remove users who return after a ban.

Our community guidelines[2] set out what sort of behaviour will and won't be accepted in our comment threads, and a user facing FAQ[3] sets out how we manage moderation.

While there is always a chance that a small number of comments are missed, even with these self-regulatory steps in place, we hope and believe that this blend of technology investment and human moderation has delivered a much enhanced user experience below the line.

April 2021

[2] https://www.theguardian.com/community-standards

[3] https://www.theguardian.com/community-faqs