

Removing social media hate speech within 24 hours sounds like a good idea, but...

Written by Fiona R Martin, Senior Lecturer in Convergent and Online Media, University of Sydney

In a bid to fight escalating anti-migrant propaganda, the European Commission this month released a [blueprint](#) for regulating online hate, which requires social media companies to take down racist material within 24 hours.

This joint [code of conduct](#) sounds like a positive political compromise. But it's unclear how it will work in practice and how it will benefit the rest of the world's social media users.

The agreement follows heavy pressure from [French](#) and [German](#) governments for Facebook to pull down racist posts, which have intensified following the recent refugee crisis.

German lawyers even started [legal action](#) against Facebook CEO Mark Zuckerberg and German manager Martin Ott, over the company's failure to remove pages sporting Nazi imagery and calling for violence against migrants.

While those [suits failed](#), Facebook, YouTube, Twitter and Microsoft have agreed to assess official reports of hate speech and "remove or disable" any that breach EU law.

The problem is that the code they signed has no concrete detail on how people can report violations, what evidence they need and how long it should take to see some action. Nor are there any accountability measures to make sure the social media giants meet the spirit of the agreement.

These are critical measures if the arrangement is to be more than a policy band-aid for a cankerous problem.

Governing hate speech

Under the code, social media companies must respond to "valid" reports of illegal hate speech as defined under the [EU's 2008 framework](#)

Removing social media hate speech within 24 hours sounds like a good idea, but...

Written by Fiona R Martin, Senior Lecturer in Convergent and Online Media, University of Sydney

But it seems Facebook and the other platforms will get to determine what qualifies as valid, unless the report comes via a civil society organisation (CSO) or a trusted EU state reporter.

That means the EU needs to nominate bodies to act as expert hate filters, to deal with public complaints, weed out vexatious claims and subtle racism, and advocate for the marginalised.

In turn, social media platforms have pledged to post CSO contact details and even to train them in reporting “with due respect to the need of maintaining their independence and credibility”.

So far so plausible. This is the type of public-private governance model [Europe](#) now favours for protecting digital media freedoms and spreads the job of internet policing among key stakeholders.

The problem is the lack of independent oversight. There’s no procedure yet for reporting back to the wider public about how reports were handled, what didn’t make the cut and what was taken down.

This may emerge, as the EU grinds towards a first assessment of progress by the end of the year, but it’s troubling that accountability isn’t built in to core of the agreement.

And beyond the EU?

The code is an old-fashioned political compromise in that it applies only to hate speech that’s published in the EU, by EU residents or on their behalf.

Yet territorial boundaries make little legal sense in internet content regulation. As the infamous [Dow Jones vs Gutnick](#) defamation case suggests, a hate speech injury happens when and where the offending text is viewed, regardless of where in the world it was published.

David Rolph, associate professor in law at the University of Sydney, also told me there is still a live issue as to where something is published online. “Is it where it’s uploaded or where it’s accessible?” he asked.

It’s likely then that the geographic cordon will be difficult to police, and that takedowns will inevitably affect people beyond the EU.

Talking back to the haters

In the meantime, the aspirational part of the code has CSOs launching their own anti-hate campaigns online, mobilising supporters to contradict, deride and parody the haters.

This is Facebook’s preferred approach to hate speech, based on United States Justice Louis Brandeis’ [vintage argument](#) that counter narratives can “expose through discussion the falsehood and fallacies, [and] avert the evil by the processes of education”.

But counterspeech also exposes activists to further attacks, as Finnish journalist [Jessika Aro](#) found out when she started to expose Russia’s troll army.

A recent [Demos study](#) for Facebook reveals we know little about whether counterspeech works to deter haters and we lack ways to measure its impact across social networks.

So rather than being a ready remedy for online violence this code signals yet more complexity in how hate speech will be regulated. It also makes promises about how social media and civil society can work together that deserve far more public debate.

If a workable process can be hammered out though, the code could be extended to any Western media system (at least where agencies have the resources to take on reporting). The question is, will the new media czars be willing to negotiate with governments outside Europe and under what terms?

Removing social media hate speech within 24 hours sounds like a good idea, but...

Written by Fiona R Martin, Senior Lecturer in Convergent and Online Media, University of Sydney

Fiona R Martin is an Australian Research Council funded Discovery Early Career Research award fellow. Her study Mediating the Conversation (DE130101267) is looking at best practice approaches to governing news commenting

Authors: Fiona R Martin, Senior Lecturer in Convergent and Online Media, University of Sydney

Read more <http://theconversation.com/removing-social-media-hate-speech-within-24-hours-sounds-like-a-good-idea-but-60325>