Intermediary Liability Blog
The Evidence Hub for Policymakers

Self-Regulation of Online Hate Speech: What the Evidence Tells Us

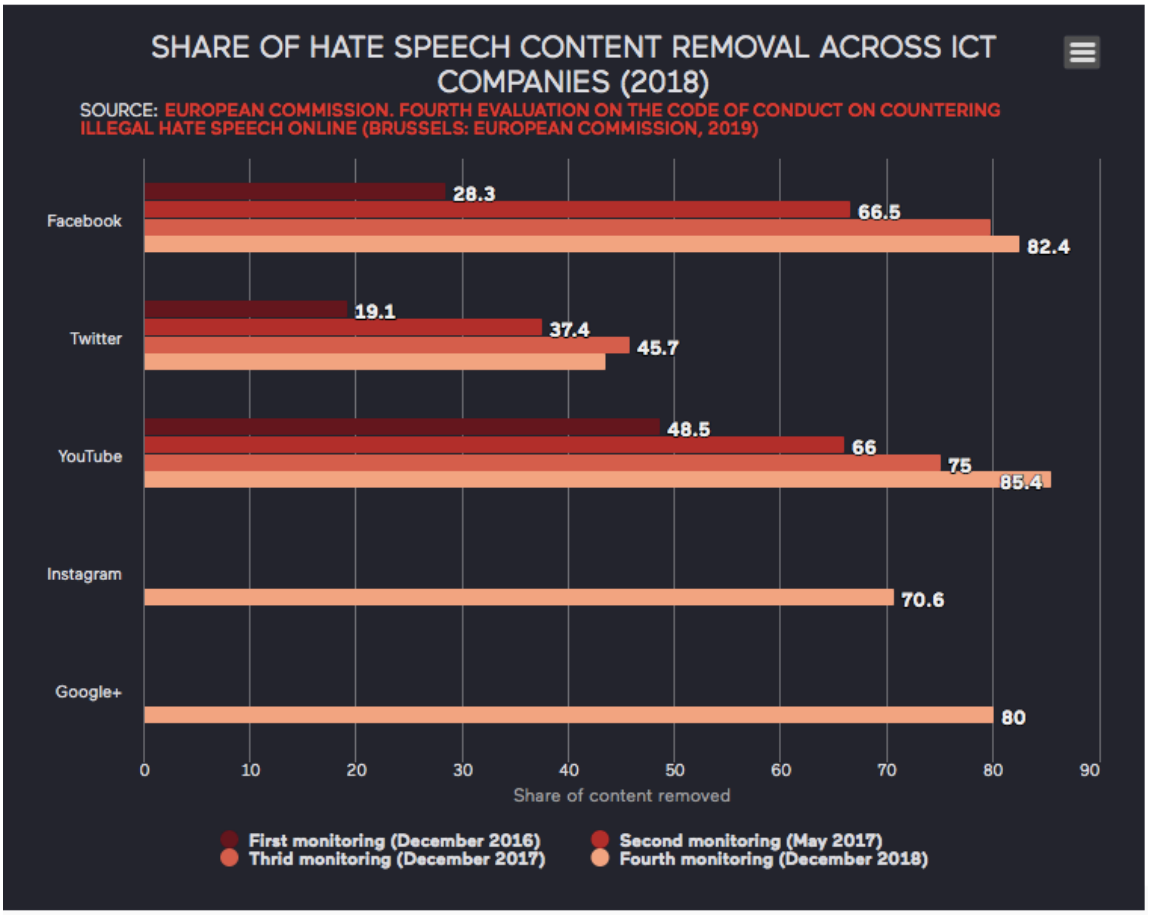
07 May 2020 | Estimated reading time: 3 minutes

Curbing hate speech is one of the most difficult challenges that regulators – and platforms – face. For starters, hate speech is fairly difficult to define. The European Commission provides a handy guide. In its Council of the European Union framework decision on combatting certain forms and expressions of racism and xenophobia by means of criminal law (2008), it calls hate speech “all conduct publicly inciting to violence or hatred directed against a group defined by reference to race, colour, religion, descent or ethnic origin, when carried out by the public dissemination or distribution of tracts, pictures or other materials.”

But how do we determine precisely which comments are or are not hate speech and what does and doesn’t need to be removed from platforms? The issue, obviously, touches directly on questions of what constitutes free speech. And, if someone is to be deciding that some speech breaks the law or violates cultural norms, who exactly should take those decisions? On what basis? And should the platforms be held accountable – both for any hate speech they miss as well as for the comments they might overzealously bring down?

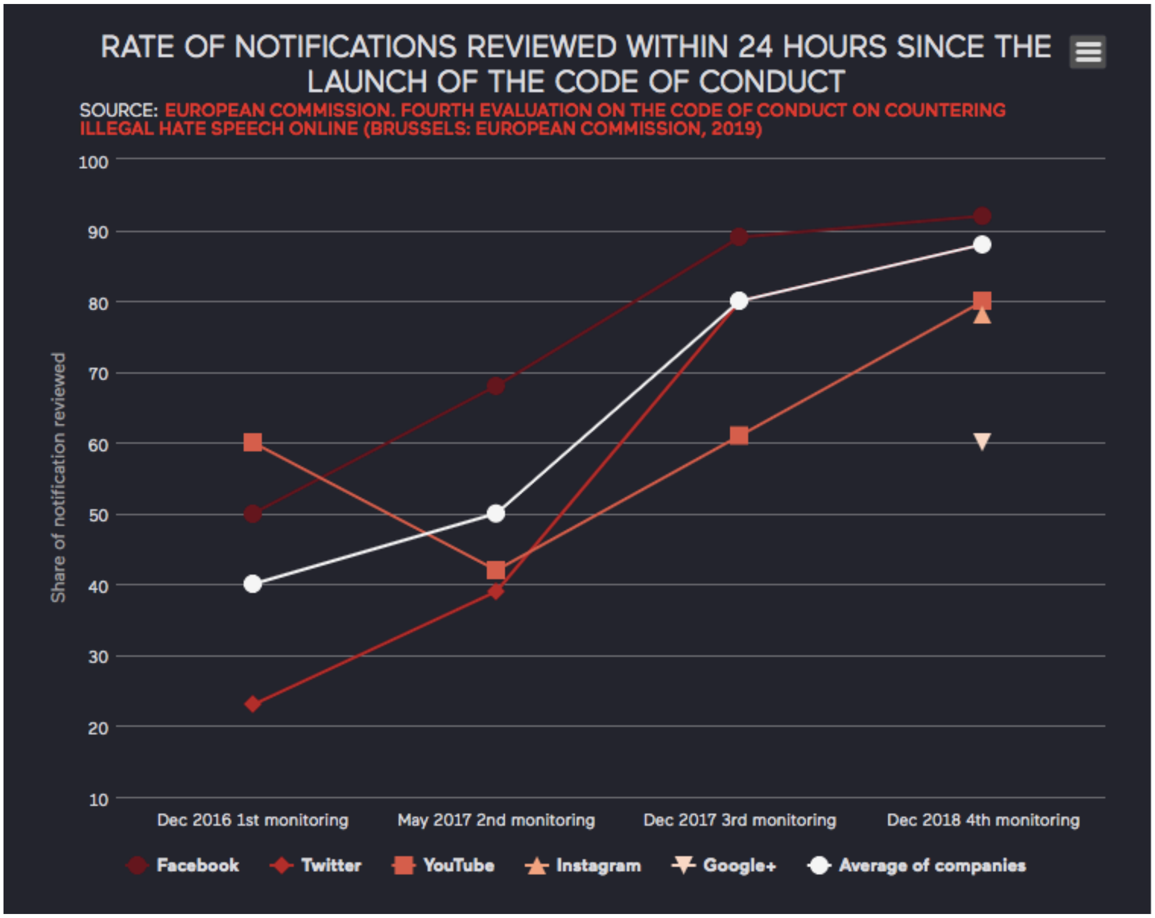
The European Commission has proposed a novel way of addressing this. Drawing industry and civil society together in a unique dialogue, it fomented a code of conduct on countering hate speech online (2016). Under this arrangement, civil-society organisations – particularly those with a background in spreading tolerance – are empowered to monitor activity on platforms. They flag content they find objectionable to the platforms, which commit to evaluate every piece of content flagged this way in under 24 hours. Each year, the NGOs file a report on how much of the content they flagged was taken down. The platforms themselves also have their own “community standards,” upon which they can ban some posts if they consider it violates their policy. And other countries – most notably Germany – have laws on what exactly constitutes hate speech and what speech is out and out illegal.

But is it working? The European Commission’s fourth evaluation on the code of conduct on countering illegal hate speech online (2019) states, broadly speaking, yes, it is. The code and its built-in monitoring mechanism “proves to be an effective tool to face this challenge,” the European Commission concludes, adding that the evidence compiled “confirms self-regulation works.”


SHARE OF HATE SPEECH CONTENT REMOVAL ACROSS ICT COMPANIES (2018)

SOURCE: EUROPEAN COMMISSION. FOURTH EVALUATION ON THE CODE OF CONDUCT ON COUNTERING ILLEGAL HATE SPEECH ONLINE (BRUSSELS: EUROPEAN COMMISSION, 2019)

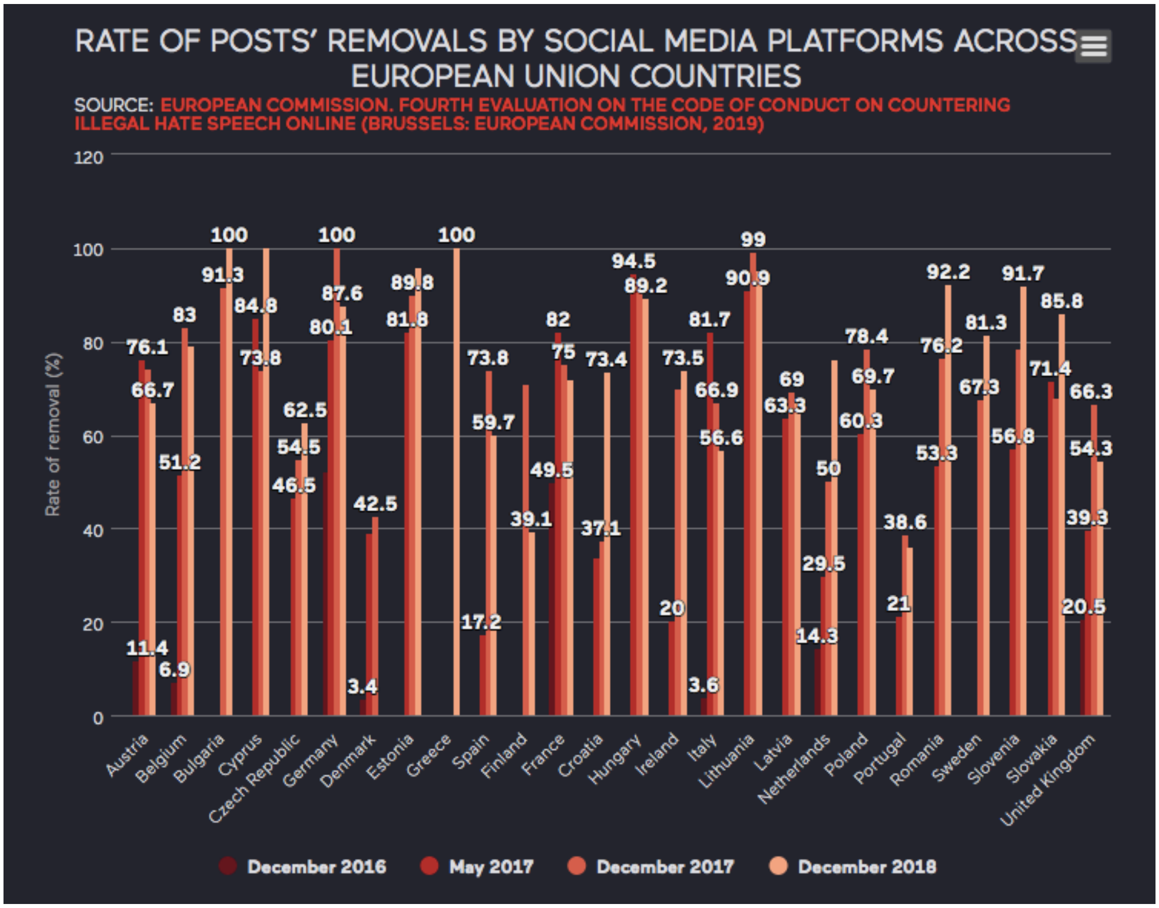
The evidence also shows that platforms have been moving quickly to respond. The European Commission’s most recent assessment found that 71.7% of the content flagged was removed – much of it in under 24 hours (88.9%). All of the global platforms showed a more aggressive stance towards banning questionable content since the advent of the code of conduct. Facebook, for one, saw its take-down rate sore to 82.4% of flagged content in 2018, up from 28.3% in 2016. But takedown rates among the major platforms also varied. All showed improvement; but YouTube was the most aggressive, with 85.4% of flagged content removed in 2018. Twitter removed the least – only 43.5% in 2018, up from 19.1% in 2016.



RATE OF NOTIFICATIONS REVIEWED WITHIN 24 HOURS SINCE THE LAUNCH OF THE CODE OF CONDUCT

SOURCE: EUROPEAN COMMISSION. FOURTH EVALUATION ON THE CODE OF CONDUCT ON COUNTERING ILLEGAL HATE SPEECH ONLINE (BRUSSELS: EUROPEAN COMMISSION, 2019)

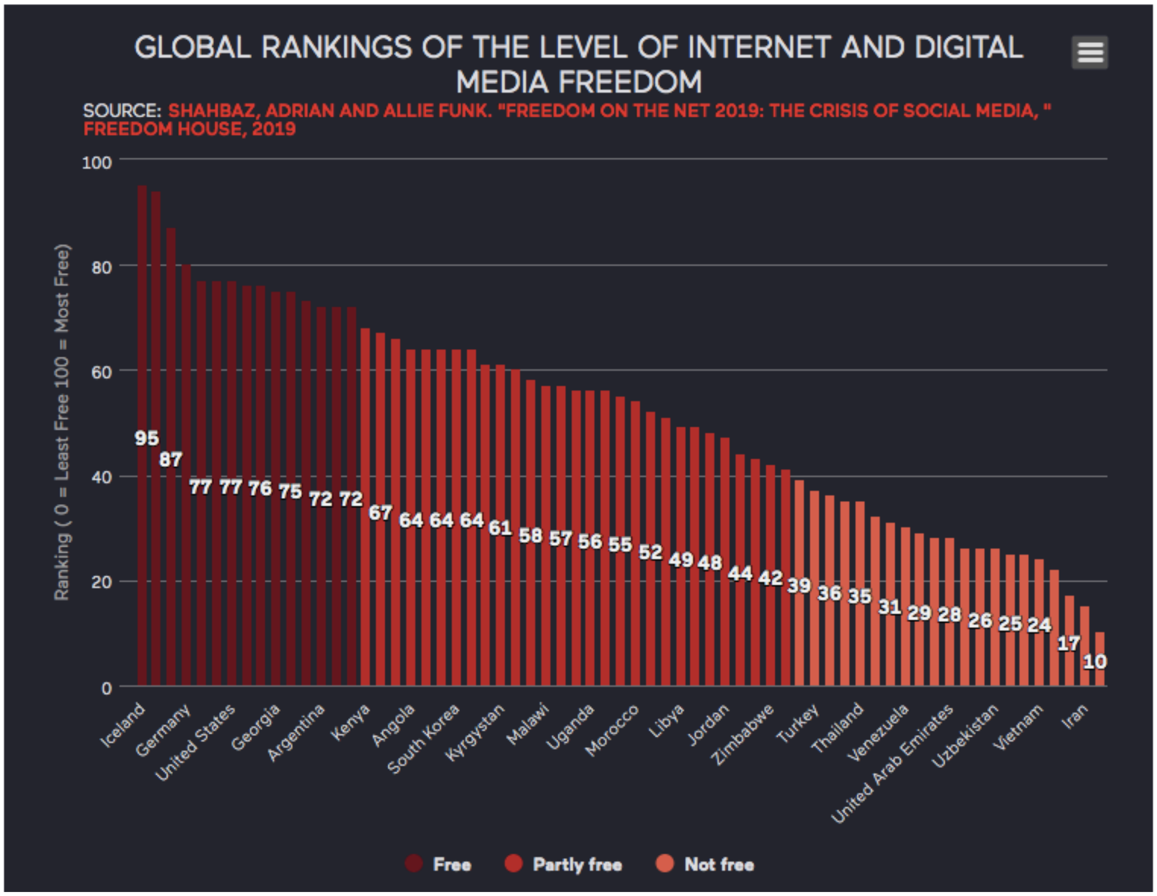
The biggest differences are found not in the comparative rates between platforms but in the muscular way that some countries approach hate speech – and are willing to accept curbs on free speech to enforce it. Germany, for one, has very strict laws banning political hate speech of all types. Its 100% removal rate for flagged content in 2017 reflects the tough law in which the platforms must operate there, including the Netzwerkdurchsetzungsgesetz (NetzDG) (2017). Other countries – such as Denmark and the United Kingdom – have looser laws and more open traditions. Platforms still respond to calls to remove more and more content in both places. But, facing looser legal requirements and more liberal traditions, the removal rate for flagged content is 42.5% and 66.3%, respectively.



RATE OF POSTS' REMOVALS BY SOCIAL MEDIA PLATFORMS ACROSS EUROPEAN UNION COUNTRIES

SOURCE: EUROPEAN COMMISSION. FOURTH EVALUATION ON THE CODE OF CONDUCT ON COUNTERING ILLEGAL HATE SPEECH ONLINE (BRUSSELS: EUROPEAN COMMISSION, 2019)

Core questions remain. For starters, how much hate speech is getting through the system? It would help to see a proper study of that. And are the trade-offs worth it? For sure, Germany has a higher take-down rate than most; but is its democracy any less rich because of it? Germany still rates a top-tercile score on Internet freedom in the Freedom House Freedom of the Net 2019 report, which measures overall obstacles to access, limits on content and violation of users' rights.



One thing is for sure: the situation with hate speech online is improving. But is it improving quickly enough? And if a nation's laws or the rights of individuals are not being violated, do we really want private companies making decisions about what goes up and what comes down online? And if so, how?

VIORICA SPAC
Viorica Spac is project manager and research associate at the Lisbon Council. She sits on the core team curating the Intermediary Liability Evidence Hub.

This blog post appeared on the Intermediary Liability Evidence Hub, an interactive website managed by The Lisbon Council, a Brussels-based think tank, to gather available evidence and data points on the issue of intermediary liability. Its website is https://evidencehub.net/.