facebook

Facebook Ireland submission to the Public Consultation on Hate Speech

Department of Justice and Equality

December 2019

Dualta Ó Broin | Head of Public Policy, Ireland

facebook

Facebook welcomes the opportunity to make a contribution to the Department of Justice and Equality's public consultation on Hate Speech. This is a matter which Facebook takes very seriously and we are eager to play a constructive role in this process by submitting our experiences and views.

Facebook echoes the comments of Minister Flanagan on launching this consultation, that the subjection of others to abuse and attack resulting from prejudice and intolerance which "can take place anywhere – on the street, on public transport, on the sports field, online and everywhere in between" is "not acceptable to the people of Ireland." As an employer with over 100 nationalities working in our offices in Ireland, Facebook is acutely aware of how important it is to create an environment of tolerance and acceptance; devoid of prejudice and intolerance.

In this submission, we aim to share information on Facebook's approach to dealing with content relating to hate speech on our platforms, which has been developed over the past 15 years, and our views on the questions posed by the consultation. Facebook notes the consultation paper's reference to the work currently undertaken by the Department of Communications, Climate Action and Environment in the regulation of online content, and how it will complement the work arising from this consultation.

Introduction

Facebook's mission is to give people the power to build community and bring the world closer together. We recognise how important it is for Facebook to be a place where people feel empowered to communicate, and we take our role in keeping abuse off our platforms seriously.

In tandem with empowering people to use their voice and express themselves, there is a need to ensure this empowerment is conducted in a safe and secure environment for all. That means we must make decisions about what is and is not acceptable amongst our diverse community of 2.8 billion people around the globe. We have therefore developed comprehensive Community Standards - a set of policies which govern what content is and is not allowed on Facebook. Our Community Standards cover areas such as hate speech; bullying; harassment; nudity; privacy and graphic violence. In developing these standards, we work with hundreds of civil society organisations and academics from around the world.

In order to enforce these Community Standards at scale, users are enabled to report content for our teams to review. Every single piece of content on Facebook - be it a photo, a status update, a comment, a profile or a page - can be reported to us for violating our policies. If the content is found to be against our Community Standards, it is removed. Facebook's specially trained content reviewers work 24/7 and support over 50 languages. These teams are part of the 35,000 people working globally on safety and security.

Policies against Hate Speech

Facebook has specific policies which set out in a clear and transparent manner that hate speech is not allowed on our platforms. Our policies on Hate Speech can be found on our Community Standards website. We define hate speech as violent or dehumanising speech, statements of inferiority, calls for exclusion or segregation based on protected characteristics, or slurs. These characteristics include race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disability or disease. In many jurisdictions our hate speech rules go beyond local law, displaying the seriousness with which we take hate speech.

We do not devise these policies alone, however. Like all other policy areas, internal expert teams consult with external experts to ensure that we consider all facets, nuances and the wider context. We are constantly evolving and changing our policies on hate speech and we continue to do so in consultation with those experts across the globe.

Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others. In some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. We permit this type of content when the intent is clear.

Enforcing our hate speech policies

During the period July-September 2019, Facebook took action on 7 million pieces of content relating to hate speech on our platforms. This is in comparison to 4.4. million pieces of content actioned during April-June 2019.

Over the last two years, we've invested in proactive detection of hate speech so that we can identify this harmful content before people report it to us and in some cases, before anyone sees it. Our detection techniques include text and image matching, which means we're identifying images and identical strings of text that have already been removed as hate speech, and machine-learning classifiers that look at things like language, as well as the reactions and comments to a post, to assess how closely it matches common phrases, patterns and attacks that we've seen previously in content that violates our policies against hate.

Initially, we've used these systems to proactively detect potential hate speech violations and send them to our content review teams since people can better assess context where AI cannot. Starting in the second quarter of 2019, thanks to continued progress in our systems' abilities to correctly detect violations, we began removing some posts automatically, but only when content is either identical or near-identical to text or images previously removed by our content review team as violating our policies, or where content very closely matches common attacks that violate our policies. We only do this in select instances.

In all other cases when our systems proactively detect potential hate speech, the content is still sent to our review teams to make a final determination. With these evolutions in our detection systems, our proactive rate has climbed to 80%, from 68% in our last report, and we've increased the volume of content we find and remove for violating our hate speech policies. During the period July-September 2019, 80.2% of actioned content was found and flagged by our technologies before users reported it.

Facebook has a "trusted partner" channel in place which allows certain NGOs or civil society organisations to report content for review. Where NGOs have a particular expertise then it can be very beneficial to allow them to have a direct reporting line to internal Facebook teams. In developing this reporting relationship over time, it facilitates the sharing of information between Facebook and trusted partners which in turn informs our ever-evolving community standards. In Ireland, we work with local NGOs that flag potentially violating hate speech content to us.

In addition, we support counterspeech initiatives across the globe by working closely with local communities, experts in civil society and academia, and policymakers. Facebook founded the Online Civil Courage Initiative (OCCI) in Europe, challenging hate speech and extremism online by partnering with local organisations to carry out training and research.

Facebook was among the first group of companies to become a signatory of the European Commission's Code of Conduct on countering illegal hate speech online in May 2016. In the most recent test, where NGOs from all over the EU reported hate speech content over a 6 week period, Facebook assessed the notifications in less than 24 hours in 92.6% of the cases. We took action on 82.4% of content that was reported to us.

As was stated by the European Commission after the fourth monitoring exercise which took place in early 2019, "Removal rates varied depending on the severity of hateful content. On average, 85.5% of content calling for murder or violence against specific groups was removed, while content using defamatory words or pictures to name certain groups was removed in 58.5 % of the cases. This suggest that the reviewers assess the content scrupulously and with full regard to protected speech."

Regulating Hate Speech Online

As our Chief Executive Officer has made clear, we welcome regulation on all forms of harmful content, including hate speech. Facebook welcomes governments and regulators around the world taking a more active role in addressing harmful content online. Protecting the people who use our services is a top priority, to which we continue to dedicate a great deal of time and resources. We do not believe any company should tackle these issues alone. This is why we work together with governments, civil society, experts and industry peers to develop rules for the internet that encourage innovation and allow people the freedom to express themselves, while protecting society from broader harms.

We regularly publish a transparency report which sets out how effectively we remove harmful content. We believe that a more standardised approach in how we and other companies do this would be beneficial. Regulation could set baselines for what is prohibited and require companies to build systems for keeping harmful content to a bare minimum.

We believe that regulation which is joined-up, consultative and collaborative in its approach is an important part of a full system of content governance and enforcement. We note Minister Flanagan's comment that, "As legislators we will also have a responsibility to strike the appropriate balance between ensuring legitimate freedom of expression and tackling unacceptable or criminal behaviour that can have devastating consequences for victims". We know from our experience that finding this balance can be hugely challenging. The provision of clarity in this nuanced area would be welcome.

Questions posed in the Consultation Document relating to The Prohibition of Incitement to Hatred Act 1989

1. **Are there other groups in society with shared identity characteristics, for example disability, gender identity, or others, who are vulnerable to having hatred stirred up against them and should be included in the list of protected characteristics?**

Facebook notes the Act's current list of protected characteristics includes "race, colour, nationality, religion, ethnic or national origins, membership of the travelling community or sexual orientation." We would suggest, at a minimum, the inclusion of age, sex, and disability on this list to bring the definition in line with the most recent working hate crime definition adopted by An Garda Síochána: "Any criminal offence which is perceived by the victim or any other person to, in whole or in part, be motivated by hostility or prejudice, based on actual or perceived age, disability, race, colour, nationality, ethnicity, religion, sexual orientation or gender." Age, gender, and disability are also already acknowledged as grounds for discrimination under Irish law by, for example, the Employment Equality Acts 1998–2015. (We note that these laws recognise "gender" as a protected characteristic but seem to attribute the meaning of "sex" to that term).

As a further step, Facebook would suggest that the protected characteristics be brought up to date and in line with society's current views and expectations. As such, we suggest also including caste, gender, gender identity, and serious disease, similar to what we have included in our Community Standards. Definitions would be required to avoid confusion about the protected characteristics' meaning.

Any suggestions for change must be made in view of the fact that an offense under the Act is criminal. Therefore, should the scope of the Act be widened, we would suggest that freedom to debate and express views relating to protected characteristics be expressly protected in the legislation.

> 2. Do you think the term "hatred" is the correct term to use in the Act? If not what should it be replaced with? Would there be implications for freedom of expression?

The Act does not define "hatred" apart from prohibiting "hatred against a group of persons in the State or elsewhere on account of their race, colour, nationality, religion, ethnic or national origins, membership of the travelling community or sexual orientation." A clear definition of hatred would be welcome to provide clarity.

Facebook notes the suggestion in the consultation document of "prejudice" or "hostility." Hostility may be interpreted as mere unfriendliness and could be argued to set too low a threshold for a criminal offense. Prejudice would be difficult to measure and define in the criminal context.

> 3. Bearing in mind that the Act is designed only to deal with hate speech which is sufficiently serious to be dealt with as a criminal matter (rather than by other measures), do you think the wording of the Act should be changed to make prosecutions under for incitement to hatred online more effective? What, in your view, should those changes be?

Although Facebook's view is that online communications currently fall within the Act's scope, any amendment to explicitly bring online communications under the Act should include a carve out so that intermediaries would not be subject to criminal liability for third-party content that violates the Act. This carve out would be particularly important if the intent requirement is amended in such a way that simply distributing or displaying violating content would be considered an offence, as intermediaries could then be held liable for content over which they have no control.

> Question 4: In your view, does the requirement that an offence must be intended or likely to stir up hatred make the legislation less effective? *and*
>
> Question 5: If so, what changes would you suggest to this element of the 1989 Act (without broadening the scope of the Act beyond incitement)?

In Facebook's view, the difficulties that arise from the requirement to prove intent or likelihood of stirring up hatred are not inherent to the requirement itself but rather stem from the lack of clarity around the definitions of both "hatred" and "stir up."

Currently, "hatred" must be directed against a group possessing one of the protected characteristics while "stir up" is generally understood to require a statement to effectively encourage others to take action against such a group. The result is that statements that express hatred towards a group without calling for specific action or that are targeted only to individuals who are part of a protected group rather than the group itself do not fall within the Act. For example, under the Act, stating "I hope that gays are beaten" would not come clearly within the definition of stirring up hatred

under the Act. Instead, the content would have to go so far as to say, “Let’s go to the local town tomorrow and beat up gays.”

Facebook would suggest an inclusion of individuals who are members of a protected group within the definition of “hatred” and a clarification that “stir up” does not require incitement to a particular act. This would therefore allow more abusive content to fall within the Act’s remit while still protecting free expression.

Conclusion

Facebook has a zero-tolerance approach when it comes to hate speech on our platforms because it creates an environment of intimidation and exclusion, and in some cases, may promote real-world violence. As such, it is firmly opposed to our mission of building communities and bringing the world closer together.

At Facebook, we believe that we have been, we are, and we will continue to be central to solving complex challenges relating to online content, including hate speech. It is clear, however, that we cannot and ought not do so alone.

We look forward to working with the Department of Justice and Equality in future stages of this important consultation process.