April 28, 2017

FOSI Briefs

Family Online Safety Institute

Technical Solutions to Controversial Content Online

It's not fake news to say that media reports have devoted much airtime and column inches to stories relating to offensive online content recently. In fact, over the last 12 month we have witnessed a considerable increase in media reports of online harassment, revenge porn, extremist videos and fake news.

We have heard requests for technology companies to "do more" about these issues. There have been calls for legislation to combat offensive behaviors, and pleas for Internet users to take responsibility for their own online experiences and to protect themselves and others. In a series of three FOSI Briefs we will examine what companies, policymakers, and everyday Internet users can do to respond to challenging content and help ensure that the online world remains as safe as possible, especially for children.

This FOSI Brief will focus specifically on the role that technology companies can play in responding to the challenges presented by controversial content. Importantly, material that is illegal is not within the purview of this article. Online child sexual abuse images, incitement to violence or terrorism, and threats of death or serious injury are clearly against the law and require a law enforcement response. The Internet industry has assisted with detection, investigation, and prosecution of offenders. In fact, tools such as Microsoft's PhotoDNA, which assigns hash values to images to help eliminate child sexual abuse images, has proven to be invaluable to efforts to eradicate illegal child sexual abuse material around the world.

In this article the problem of legal, but objectionable, content will be addressed. Primarily, this is material that the majority of people would find to be distasteful and sometimes offensive, but in itself it does not contravene any national laws. Technology companies, specifically social media platforms, by their very nature promote the sharing of information, views, and experiences, but this can also lead to challenges.

Offensive words, violent threats, exposure of private information, repeated unwanted contact, misuse of reporting tools, and denial of service attacks are all forms of online harassment. The Data & Society Research Institute and the Center for Innovative Public Health Research found that 47% of adult Internet users had been victims of harassment on the Internet. While also finding young people, women, and lesbian, gay, and bisexual Americans were more likely to be targeted. Among the more serious implications of online harassment, erosion of trust in the service and self-censorship are common reactions. For these reasons technology companies, especially social media platforms, are working hard to eradicate this harmful behavior which may impact their business model, as well as have wider societal implications.

HARASSMENT

Twitter has struggled with the issue of harassment, partly due to the fact that it allows for anonymity, as well as its public nature. Consequently, the company has been working diligently to respond to the problem while respecting its core First Amendment values. The platform relies heavily on users to report harassment when they see it, whether it occurs to them or not. The company has worked to remove any barriers to reporting and speed up response times. User experience has also been improved through increased control, including muting offensive accounts, managing repeat notifications, and preventing individuals from creating multiple profiles for harassing purposes.

Additionally, Twitter is relying on technology to stop abuse, deploying algorithms to identify accounts that may be engaged in harassment before individuals report them. These offending accounts are then limited for a certain period of time. Repeated abuse may result in permanent suspension. These are just two ways that technology companies are trying to ensure that their users have safer online experiences.

REVENGE PORN

The nonconsensual sharing of intimate images, or revenge porn, is another form of harassment for which bad actors have used social media. Facebook has responded to this abhorrent behavior by making its reporting easier and ensuring that specialist teams are available to respond to the reports, take down the image, and disable the offending account. Partner organizations are also available to offer support. Through the use of photo-matching technology, further attempts to share the image are prevented across the Facebook platforms.

The posting of intimate images without consent contravenes the terms of service of the majority of social media sites. Google, Twitter, and Microsoft have worked hard to simplify the process of reporting and removing these images. Google will remove pictures from their search results when they receive a report from the victim, while Twitter prohibits the posting of intimate images without consent in their community rules. Microsoft has an easy to use web form and commits to removing revenge porn from Bing, OneDrive, and Xbox Live.

EXTREMISM

Arguably the most controversial content online is that which relates to extremism and radicalization. FOSI has written previously about the risks posed to young people online by the new phenomenon of terrorists using the Internet to recruit new members. But recruitment is not the only problem. The use of social media and video platforms to further the objectives of terrorist's ideology is particularly concerning. However, there is a balance to be struck. What may appear to be propaganda at first sight, may actually have significant newsworthy value, or form part of a counter-narrative messaging program and as such automated removal of this material is particularly difficult.

YouTube is using a combination of notice-and-takedown, following public reports, and leveraging artificial intelligence and machine learning to ensure that highly controversial content with no significant value is removed from its services. Additionally, YouTube is promoting counter-extremist videos that are also available on YouTube (see below for more details). Twitter has also struggled with the exploitation of its service by extremists. Hundreds of thousands of accounts have been suspended as a result of user reporting, as well as leveraging existing technology such as spam-fighting tools. Both platforms continue to explore new ways, technological and otherwise, to ensure that terrorists are not misusing their services.

A truly technical response to controversial online content has been developed by Jigsaw, an incubator within Alphabet. The team of engineers, researchers, and designers have developed new tools to combat radicalization and online harassment. Using existing technology and content, "The Redirect Method" diverts those looking for extremist content to counter-narrative material that is already available on the Internet. This approach targets those specifically at risk and harnesses the power of advertising alongside keywords and phrases to counter radical messaging. Meanwhile, "Perspective" uses machine learning to target online abuse and harassment. It assists moderators in sorting comments by scoring words and phrases based on the potential impact that they may have on a conversation. This combines a technological response with human review and input.

New technology promises much to combat the spread of controversial content online, but the nuances that come with speech will always require some level of human involvement. Whether that comes from formal moderation, or community reporting, everyone has a role to play. It has never been so important to teach children about their online rights and responsibilities. All users have the right to use the Internet free from harassment, fear, and confusion, but in order to achieve that the community must come together. This includes technology companies, policy makers, parents and children. There remains considerable hope for a civil Internet in the future, but all must take responsibility not only for their own online experience, but also for that of others.

Emma Morris
Global Policy Manager
emorris@fosi.org

About FOSI

The Family Online Safety Institute is an international, non-profit organization which works to make the online world safer for kids and their families. FOSI convenes leaders in industry, government and the non-profit sectors to collaborate and innovate new solutions and policies in the field of online safety. Through research, resources, events and special projects, FOSI promotes a culture of responsibility online and encourages a sense of digital citizenship for all.