Big Tech is overselling AI as the solution to online extremism

September 17 2018, by Kyle Matthews And Nicolai Pogadl

Credit: CC0 Public Domain

In mid-September the European Union threatened to fine the Big Tech companies if they did not remove terrorist content within one hour of appearing online. The change came because rising tensions are now developing and being played out on social media platforms.

Social conflicts that once built up in backroom meetings and came to a head on city streets, are now building momentum on social media platforms before spilling over into real life. In the past, governments tended to control traditional media, with little to no possibility for individuals to broadcast hate.

The digital revolution has altered everything.

Terrorist organizations, most notably Islamic State (ISIS) militants, have used social media platforms such as Facebook, Instagram and Twitter for their propaganda campaigns, and to plan terrorist attacks against civilians.

Far right groups, including anti-refugee extremists in Germany, are also increasingly exploiting tech platforms to espouse anti-immigrant views and demonize minorities.

From Sri Lanka to Myanmar, communal tensions —stoked online —have led to violence.

Due to the growing political will within Western countries to regulate social media companies, many tech titans are arguing they can self-regulate—and that artificial intelligence (AI) is one of the key tools to curtail online hate. Several years ago, we created the Digital Mass Atrocity Prevention Lab to work on improving public policy to curb the exploitation of tech platforms by violent extremists.

Oversold abilities?

Tech companies are painfully aware of the malicious use of their platforms.

In June 2017, Facebook, Microsoft, Twitter and YouTube announced the formation of the Global Internet Forum to Counter Terrorism, which aims to disrupt extremist activities online. Yet as political pressure to remove harmful online content grows, these companies are beginning to realize the limits of their human content moderators.

Instead, they are increasingly developing and deploying AI technologies to automate the process of unwanted content detection and removal. But they are doing so with no oversight and little public information about how these AI systems work, a problem identified in a recent report by the Public Policy Forum.

Twitter, according to its most recent transparency report, claims it used AI to take down more than 300,000 terrorist-related accounts in the first half of 2017.

Facebook itself acknowledges that it is struggling to use make use of AI in an efficient manner on issues surrounding hate speech. CEO Mark Zuckerberg told members of the U.S. Congress earlier this year that AI still struggles to tackle the nuances of language dialects, context and whether or not a statement qualified as hate speech —and that it could take years to solve.

However, the company also claims to be able to remove 99 per cent of ISIS and al-Qaida affiliated content using AI-powered algorithms and human content moderators. Whether AI or humans are the key to the company's claims of success has not yet been independently investigated.

The failure of AI

In 2017, 250 companies suspended advertising contracts with Google over its alleged failure to moderate YouTube's extremist content. A year later, Google's senior vice president of advertising and commerce, Sridhar Ramaswamy, says the company is making strong progress in platform safety to regain the lost confidence of its clients.

However, a recent study by the NGO Counter Extremism Project refutes the effectiveness of the company's effort to limit and delete extremist videos. More transparency and accountability from YouTube is needed, given that the study found that over 90 per cent of ISIS videos were uploaded more than once, with no action taken against the accounts that violated the company's terms of service.

Clearly there is no simple pathway forward. Removing content that is not harmful, offensive, extremist or illegal, even if it distasteful, is an impediment to free speech. In some cases, using AI to remove content has blocked legitimate material posted by human rights champions.

For example, in 2017, Shah Hossain, a human rights activist found a significant number of his Facebook posts regarding the persecution of the Rohingya minority in Myanmar had been deleted. YouTube also erased his news channel, which had nearly 80,000 subscribers. Hossain was documenting human rights abuses, not espousing hate.

In Syria, where independent journalism is severely restricted by war, videos and photos posted online by activists are crucial to understanding the situation in the country. In an attempt to crackdown on extremist content, however, YouTube's AI-powered algorithms removed thousands of videos of atrocities against civilians. The videos were posted as evidence for the eventual prosecution of Syrian officials for crimes against humanity. This is quite troubling.

Moving forward

Well-known social media giants have said publicly that they'll put more resources into policing their platforms. However, given the current results, it's time to consider if this approach is ethical and effective.

The United Kingdom, France, Germany, the European Union and the United States, among others, have begun to openly discuss and implement regulatory measures on the tech industry, not only pertaining to terrorism and hate speech, but also digital election interference, the spread of "fake news" and misinformation campaigns.

Canada has begun to take the issue seriously as well, forming the Digital Inclusion Lab at Global Affairs Canada, which works to strengthen the combined efforts of the G7.

These are much needed initiatives. The big tech giants have been overselling the effectiveness of AI in countering hate on their platforms. Our democratic and open societies must put aside the notion that AI is the panacea for the problem at hand. Social polarization and growing mistrust across the planet will continue unless elected officials regulate Big Tech.

Provided by The Conversation

Citation: Big Tech is overselling AI as the solution to online extremism (2018, September 17) retrieved 24 June 2024 from https://phys.org/news/2018-09-big-tech-overselling-ai-solution.html