Written Representation 17

Name: Hany Farid
Professor & Chair
Computer Science
Dartmouth College

Received: 5 Feb 2018

Dear Select Committee,

I am a Professor and Chairman of Computer Science at Dartmouth College and a Senior Advisor to the Counter Extremism Project (CEP). The CEP, run by Mark Wallace, former US Ambassador to the UN and Fran Townsend, former US Homeland Security Advisor, is a not-for-profit, bipartisan, international policy organization formed to combat the growing threatfrom extremist ideologies.

Overview

The Internet promised to democratize access to knowledge, spread great ideas, and promote tolerance and understanding around the globe. This promise, however, is being poisoned by the rise of trolling, cyber-bullying, revenge porn, fake news, child exploitation,hate, intolerance, and extremism.

There is no question that reining in online abuses is challenging. There is also no question that we can and must do more than we are to mitigate the harm that is being seeded and fueled online, while at the same time maintaining an open and free Internet where ideas can be shared and debated. I reject the claim by some that these two goals are mutually exclusive.

I will begin by describing technology that I have previously developed and deployed to combat on-line child exploitation and extremism. I will also describe the limitations of this technology and the strengths and limitations of the newest advances artificial intelligence and digital forensics.

Combating Child Exploitation

Since early 2000, we have seen a pattern of denial and inaction from technology companies when it comes to responding to misuse on their platforms.

In 2003, the rise of the Internet brought an explosion in the global distribution of child pornography. United States Attorney General Ashcroft called the major US-based technology companies of the time together to discuss the problem. This meeting eventually led to the formation of the Technology Coalition with a stated mission of “eradicating online child sexual exploitation.” Between 2003 and 2008, however, the Technology Coalition did not develop or deploy any effective solution to disrupt the global distribution of child pornography—the problem, in the intervening five years, only worsened.

Because of my background and expertise in digital forensics, I was approached in 2008 by Microsoft and the National Center for Missing and Exploited Children (NCMEC) and asked to try to understand why a solution was so elusive and what, if any, technology could be developed and deployed to help reduce the global distribution of child pornography. In response, we developed and deployed a technology called PhotoDNA that is now in worldwide use and has been effective in removing tens of millions of images of child exploitation from online platforms.

Combating Extremism

It is from my work and experience with fighting online child exploitation that I partnered with the CEP to address the troubling online recruitment and radicalization of extremists and the resulting devastating consequences in the form of attacks around the world.

The horrific aftermath of extremists' misuse of the Internet and social media platforms stretches from Paris, to Brussels, to London, Orlando, San Bernardino, Istanbul, Beirut,
Cairo, and beyond. Since 2014, CEP, along with government agencies around the world, has been calling on technology companies, and social media in particular, to rein in the ability of extremists to recruit, radicalize, plan, and execute attacks. The response—largely denial and inaction—has been eerily similar to the call in early 2000 to combat online child exploitation.

Finding and removing extremism-related material is a difficult problem, but as with child exploitation and copyright infringement material, there are effective technological solutions that can be deployed. Building on the success of PhotoDNA, we at CEP have developed the next generation of technology that can accurately and quickly find and remove known extremist-related material in the form of digital images, videos, or audio recordings. Unveiled in June 2016, this technology, which we call eGlyph, extracts from digital content a signature that is both distinct (two different images/videos don't share the same signature) and is stable as the content is either intentionally or unintentionally modified as it makes its way around the Internet. In addition, we continue to develop and deploy new technologies that we think should be widely deployed to remove content that is clearly designed to recruit and radicalize extremists.

Content Moderation

As described above, technology exists to accurately and efficiently find and remove content *once it has been identified by human moderators*. This technology extracts from a digital image, audio, or video a distinct and stable signature which can then be compared to all future uploads. The strength of this technology is that it can quickly and accurately identify content and has a proven track record of working on some of the world's largest platforms (Facebook, Google, Microsoft, etc.).

The limitation of this technology is that it cannot, without human assistance, discover new illegal or inappropriate content. Another limitation is that this technology is applicable only to audio, image, and video recordings but is not applicable to text-based content.

Content Classification and Authentication

Although the past few years have seen dramatic improvements in artificial intelligence (AI) and machine learning (ML), this technology is not, in my opinion, ready for fully autonomous large-scale deployment. The primary issue is one of "false alarms" — mis-classifying content as inappropriate or illegal. The photoDNA and eGlyph technologies described above achieve a false-alarm rate on the order of 1 in 50 to 100 billion. Even the best AI/ML content classification technologies are operating at false alarm rates at no better than 1 in 1000. At the scale of the internet, with billions of uploads a day, this false alarm rate is prohibitive. Significant advances are still required before this relatively new technology will be ready for largescale deployment.

In addition to the content moderation technologies described above, advances in digital forensics are being developed to detect fake audio, image, or videos. As with AI/ML classification technologies, techniques in digital forensics are not yet accurate enough to operate at large-scale. The issue remains one of false-alarms that are prohibitively large.

The latest AI/ML and digital forensics technologies may, however, be effective when paired with human reviewers who can manually review any flagged content. It is most likely the case that any content moderation in the coming years will require this type of joint computerand human-based review.

Conclusions

Some technologies exist today that can be deployed to effectively reduce the impact of harmful content from child-exploitation to extremism-related material and fake news. Other newer technologies will need to be paired with manual review to overcome limitations in these technologies. Significant effort should be placed in developing and deploying new technologies, putting pressure on technology companies to do more to rein in abuses on their platforms, and in informing the public on how to more critically digest digital content.

I hope that you will find these thoughts helpful in your deliberations.

Sincerely yours,

Hany Farid, Ph.D.

References:

1. H. Farid. Reining in Online Abuses. Technology and Innovation, 2018
2. Are Internet companies complicit in promoting hateful and harmful content?, October 2017 (https://www.neweurope.eu/article/internet-companies-complicit-promoting-hateful-harmful-content/)
3. H. Farid. Photo Forensics. MIT Press, 2016
4. V. Conotter, E. Bodnari, G. Boato, and H. Farid. Physiologically-based Detection of Computer Generated Faces in Video. *International Conference on Image Processing*, Paris, France, 2014.
5. DARPA Media Forensics Program (www.darpa.mil/program/media-forensics)