The moderation of extremist content is prone to error, causing real-world harm

Jillian C. York

08.02.2022 doi: 10.17176/20220208-121024-0

In the mid-1990s, a questionable and now-debunked study from an undergraduate student about the prevalence of pornographic imagery online sparked a TIME Magazine cover story, leading to a Congressional inquiry and eventually to the creation of the Communications Decency Act (CDA). The Act included a provision that imposed sanctions on anyone who "uses an interactive computer service to send to a specific person or persons under 18 years of age or [...] uses any interactive computer service to display in a manner available to a person under 18 years of age, any comment, request, suggestion, proposal, image, or other communication that, in context, depicts or describes, in terms patently offensive as measured by contemporary community standards, sexual or excretory activities or organs."

Then-president Bill Clinton signed the Act into law in 1996, prompting opposition from civil rights groups. The American Civil Liberties Union (ACLU) filed a lawsuit, arguing that the censorship provisions of the CDA were unconstitutional because they would criminalize protected expression, as well as because the terms "indecency" and "patently offensive" were unconstitutionally vague. In a landmark 1997 ruling, the Supreme Court ruled that the CDA created an "unacceptably heavy burden on protected speech" that threatened to "torch a large segment of the Internet community." In the decision, Justice John Paul Stevens wrote that "the interest in encouraging freedom of expression in a democratic society outweighs any theoretical but unproven benefit of censorship."[1]

The CDA was overturned and pornography eventually proliferated across the internet, but the debacle nevertheless had a lasting impact, creating a bright dividing line between sexually explicit websites and the rest of the web. By the time "Web 2.0" came around, most major sites had in place anti-obscenity policies, even in the absence of most other restrictions on speech.

As such, while pornography remains prevalent across the web, there has been a broad chilling effect on sexual expression, as well as the expression of non-sexual nudity and sexual health information. Government and company policies intended to keep lascivious content away from the eyes of minors (such as Facebook's complex community standard around nudity and sex) ultimately end up casting a wide net, preventing a range of expression and information-sharing.

Old tactics, new purpose

Today, we are witnessing a similar phenomenon when it comes to extremist and terrorist content online: Policies intended to limit the ability of terrorist groups to organize, recruit, and incite—as well as for individuals to praise such groups—have been expanded in recent years and often result in the erasure of not only extremist expression, but human rights documentation, counterspeech, and art.

While some restrictions are legally required, many corporate anti-extremism policies were created in response to pressure from NGOs, governments, and the public and have little or no basis in law. For instance, while groups such as the Iranian Revolutionary Guard Corps (IRGC) are sanctioned by the United States and therefore cannot be hosted by U.S. platforms, there is no law preventing individuals from discussing the merits of their role in Iranian society.

Despite that, companies like Facebook and Google have taken a particularly hard-line approach in recent

[1] Jillian C. York, *Silicon Values: The Future of Free Speech Under Surveillance Capitalism* (Verso: 2021), p. 147.

years against such groups, relying on similar tools and tactics that have long been used to moderate sexually explicit content.

In the years following the repeal of the CDA, it was not only sexually explicit adult material that proliferated across the web, but also child sexual abuse imagery (CSAM). The need to swiftly remove such content (while causing as little harm to human content moderators as possible) required the creation of tools that could identify CSAM and automate its removal. The existence of a law enforcement database of CSAM made this relatively easy, as images that showed up online could be matched to those in an existing database. Thus, PhotoDNA — a technology that identifies CSAM and matches it to material in a database based on unique fingerprints, or hashes — was born. When images are detected, they are reported to the National Center for Missing and Exploited Children (NCMEC) as required by law.

The creator of PhotoDNA, Dr. Hany Farid — a professor at the University of California, Berkeley — suggested its use for detecting terrorist imagery, initially to little interest. But as the use of social media by groups such as the Islamic State increased in both sophistication and scale, companies began to adopt the technology, relying on a database run by the Global Internet Forum to Counter Terrorism (GIFCT). Though GIFCT member companies are not required to use the database, most do, at least in part.

The trouble with defining 'terrorism'

The race by companies to eradicate extremism is arguably more complex than the one to fight child sexual abuse imagery, however, and for a number of reasons. First, there is no globally agreed-upon definition of "terrorism," and throughout modern history, states have used the term to classify and deny rights to their opponents. A quick glance at the United States, European, and United Nations lists of terrorism organizations shows substantive differences in approach.

Second, most major social media platforms are subject to U.S. law in some way or another, whether or not they are headquartered in the United States. They must comply with certain U.S. sanctions but are also under pressure to default to U.S. classifications—classifications which are often political in nature.[2] Without an internationally agreed-upon definition of terrorism, companies—and by extension, the public—are forced to trust the GIFCT and its member companies' definition.

This has proven troublesome. All areas of content moderation include some errors, whether the moderation is undertaken by humans, artificial intelligence, or some combination of the two, and companies are not typically forthcoming about their error rates. When it comes to online extremism specifically, there is reason to believe that over-moderation is quite common, owing to the aforementioned complexities and the rigidity and typically binary nature of how content policies are enforced.

There are numerous illustrative examples that demonstrate the complexities of moderating extremist imagery. In 2017, for instance, an Emirati journalist posted a photo which featured a picture of Hezbollah leader Hassan Nasrallah with a rainbow Pride flag overlaid across it, intended as a satirical statement. Although satire is a permitted exception to the company's rules against terrorist content, the image was removed because it contained the photo of a designated terrorist.[3]

Documents leaked to the *Guardian* around the same time period demonstrated that Facebook moderators are trained to remove imagery that contains support, praise, or representation of terrorist groups, and to ignore those that are presented in a neutral or critical fashion. But human moderators must make split-second decisions, and must therefore memorize the faces of numerous designated individuals. The propensity for error is apparent.

[2] Alex Schmid, "Terrorism: The Definitional Problem," Case Western Reserve Journal of International Law, 2004: Vol. 36, Issue 2, https://scholarlycommons.law.case.edu/cgi/viewcontent.cgi?article=1400&context=jil
[3] York, *Silicon Values*, p. 387

Automation is error-prone

Automation seems even more prone to error than humans in this area. Training data libraries can create normative attributes for certain types of image classifications; for instance, a body with large breasts is assumed to belong to a woman.[4] The image of Nasrallah, then, presented without comment, would not be read as satirical if the machine learning algorithm was not trained to pick up on the overlaid rainbow flag.

Training a machine learning algorithm for the purpose of removing terrorist imagery requires the creation of a dataset that includes a significant amount of content in one category, which is then fed to the algorithm for training. For example, in order to accurately identify extremist content, a company like YouTube would create a set of data that it defines as extremist—such as a large number of ISIS videos—then feed that data to its algorithm. Any mistakes made by the algorithm can be difficult to understand—unless specifically designed to be interpretable, machine learning algorithms cannot be understood by humans.

Moderating text using machine-learning algorithms can be even more complex than moderating imagery. The Open Technology Institute describes the challenge thus: "In the case of content such as extremist content and hate speech, there are a range of nuanced variations in speech related to different groups and regions, and the context of this content can be critical in understanding whether or not it should be removed. As a result, developing comprehensive datasets for these categories of content is challenging, and developing and operationalizing a tool that can be reliably applied across different groups, regions, and sub-types of speech is also extremely difficult. In addition, the definition of what types of speech fall under these categories is much less clear." The Institute concludes that "these tools are limited in that they are unable to comprehend the nuances and contextual variations present in human speech."

It is not clear how much of the GIFCT database consists of human-identified versus machine-identified content, namely because the database has not been shared with any members of civil society focused on human rights (despite demands), and the GIFCT has offered minimal information about how the database functions. Unlike the databases used for identifying CSAM, the GIFCT database has no external oversight.

Thus, any error contained within the database is multiplied across the social web. And such errors abound: In addition to satirical content such as that in the previous example, documentation of human rights violations and violent conflict is regularly removed as well. According to Syrian Archive, a group that documents and archives such content, at least 206,077 videos documenting rights violations were made unavailable on YouTube between 2011 and May 2019.

It is no surprise that companies have undertaken such drastic measures against extremist content. Politicians, law enforcement, and other officials regularly engage in hyperbole that encourages strong action whilst ignoring the potential pitfalls. An oft-repeated quote from Commander Dean Haydon, of the Met's Counter Terrorism Command states that "Every tweet has the potential to radicalize vulnerable people," while terms like "eradicate" and "eliminate" have been used by efforts such as the Christchurch Call, an effort put forward by the governments of New Zealand and France following the attacks on a Christchurch mosque by a white supremacist in 2019.

The development of the Christchurch Call, a multi-stakeholder effort that includes governments, companies, and civil society and which works closely with the GIFCT, raises another key issue; that decisions about who is or not a terrorist, and who gets to make such a decision. The Christchurch Call was created in response to a white supremacist act of terror, and yet the focus of GIFCT—and ostensibly of the database it oversees—has been on Islamist groups such as ISIS and Al Qaeda. This demonstrates a clear disconnect in goals but also illuminates a core problem with the effort to "eradicate" terrorism: The world is not in agreement about what constitutes terrorism and, as previously mentioned, various (or perhaps most) governments have manipulated this term to suit

[4] York, *Silicon Values*, p. 388

their political goals. As such, the lack of oversight and minimal expert and civil society involvement is exceptionally troubling.

ABOUT THE AUTHOR

Jillian C. York

Jillian C. York is the the author of Silicon Values: The Future of Free Speech Under Surveillance Capitalism, the Director for International Freedom of Expression at the Electronic Frontier Foundation, and a visiting professor at the College of Europe Natolin.Jillian C. York is the the author of Silicon Values: The Future of Free Speech Under Surveillance Capitalism, the Director for International Freedom of Expression at the Electronic Frontier Foundation, and a visiting professor at the College of Europe Natolin.

SUGGESTED CITATION

York, Jillian C.: The moderation of extremist content is prone to error, causing real-world harm, Verfassungsblog, 08.02.2022, doi: 10.17176/20220208-121024-0.

Licensed under CC BY-SA 4.0.

Langzeitarchiviert bei IntRechtDok unter: doi.org/10.17176/20220208-121024-0.