The Intercept_

The White House Asked Social Media Companies to Look for Terrorists. Here's Why They'd #Fail.

Jenna McLaughlin

January 20, 2016

The White House asked internet companies during a counterterrorism summit earlier this month to consider using their technology to help "detect and measure radicalization."

"Should we explore ways to more quickly and comprehensively identify terrorist content online so that online service providers can remove it if it violates their terms of service?" asked a White House briefing document that outlined the main topics of conversation for the meeting. The document, which was obtained by *The Intercept*, is now posted online.

The briefing suggested that the algorithm Facebook uses to spot and prevent possible suicides might be a helpful model for a technology to locate terrorists, asking: "Are there other areas where online providers have used technology to identify harmful content and remove it? … Something like Facebook's suicide process flow?"

Government officials also want to use such an algorithm for law enforcement purposes. "Are there technologies that could make it harder for terrorists to use the internet … or easier for us to find them when they do?" read the briefing.

"We are interested in all options to better identify terrorist networks, or indications of impending plots. … Are there ways to glean from changes in patterns of use of these platforms involvement in preparations for violence?"

An increasingly large proportion of terrorism investigations these days start with tweets or posts, generally flagged by family members or informants. Civil libertarians worry that the FBI is using protected speech to identify potential subjects of entrapment. But the FBI's concern is that it's not seeing everything it needs to see.

And at the same time, there's increased pressure for social media companies to deny radical groups an open platform for speech.

No wonder the government wants an algorithm.

But there are some major problems with trying to use computer code to find "terrorists" or "terrorist" content.

First of all, it doesn't work. Many experts, including people with law enforcement, academic, and scientific backgrounds, agree that it's practically impossible to boil down the essential predictive markers that make up a terrorist who is willing and capable of carrying out an attack and then successfully pick him out of a crowd.

"Many believe that data mining is the crystal ball that will enable us to uncover future terrorist plots. But even in the most wildly optimistic projections, data mining isn't tenable for that purpose," wrote Bruce Schneier, prominent cryptologist and fellow at Harvard's Berkman Center for Internet and Society, in 2006.

Despite hyped-up cable news coverage and fearmongering messages from government officials, terrorism is an incredibly rare event in the United States. According to the New America Foundation's attack tracker, there have been a total of nine "violent jihadist attacks" on U.S. soil since September 11, 2001, resulting in 45 deaths.

Algorithms are good at some things — like correctly concluding that your credit card has been stolen. But that's because it happens so commonly, there are not many variables, and incidents follow a predictable pattern.

"Something as unique and rare as terrorism — that's what makes this different from credit card fraud," Schneier told *The Intercept.*

Consider medical testing, Schneier said. "When a disease is very rare, if your test tests positive, it's almost always wrong, because your chances of having that disease are one in a million."

Think about that for a minute: Imagine you're trying to determine who has that incredibly rare disease, and it can be spotted by genetic testing.

Say your test is 90 percent accurate in determining whether someone suffers from that disease. That means it is also wrong 10 percent of the time. One out of 10 of your patients will test positive, even though chances are that none of them have the disease.

Out of a million people, 10,000 would test positive — but chances are only one would really have the disease. And you wouldn't know which one.

Now imagine the odds are one in 100 million, amid many hundreds of millions of social media postings. Imagine how many posts would be deleted or referred to law enforcement in error.

And keep in mind there's no real way to come up with a test for terrorism that's even 90 percent accurate. There's not even a good statistical database of people charged for terrorism-related crimes, just for starters.

False positives when using algorithms to spot suspected credit card fraud have little cost. "A call to the customer from a credit issuer will reassure the customer whether he or she is correctly targeted or not," said Jim Harper, a senior fellow at the Cato Institute, during a Senate Judiciary Committee hearing on data mining in 2007.

But “identifying” terrorists is a different matter. “Because of the statistical impossibility of catching terrorists through data mining, and because of its high costs in investigator time, taxpayer dollars, lost privacy, and threatened liberty, I conclude that data mining does not work in the area of terrorism,” Harper said.

“Of course there’s no way for software to identify and remove terrorists or terrorist content from online media,” Phil Rogaway, a computer science professor at UC Davis, wrote in an email to *The Intercept*. “A group of humans would routinely disagree if a given email or post constitutes terrorist content, so how on earth is a program to make such a determination?”

A 2008 government study also concluded that counterterrorism data-mining programs seeking patterns in personal information, like travel records, phone records, and website browsing history, were ineffective and should be evaluated for privacy impacts.

Local “fusion centers” designed to share intelligence and data on terrorism and report back to the Department of Homeland Security were described by Senate investigators in 2012 as “oftentimes shoddy, rarely timely, sometimes endangering citizens’ civil liberties and Privacy Act protections, occasionally taken from already-published public sources, and more often than not unrelated to terrorism.”

And what if such an algorithm is put into action, and starts automatically deleting posts?

In the briefing document, the administration asks whether “technologies used for the prevention of spam” might be useful in locating and removing terrorist content.

But a filter like that could easily snatch up First Amendment protected speech.

“Censorship has never been an effective method of achieving security, and shuttering websites and suppressing online content will be as unhelpful as smashing printing presses,” said former FBI agent Michael German, a fellow at the Brennan Center for Justice.

I shared some passages from the White House briefing with him. “These passages make clear that the government continues to cling to long-disproven, simplistic theories of terrorist radicalization, which suggest that the exposure to extreme ideas leads to terrorist violence,” he wrote in an email to *The Intercept*. “If ideas are identified as the problem, the only solution can be the suppression of those expressing such sentiments.”

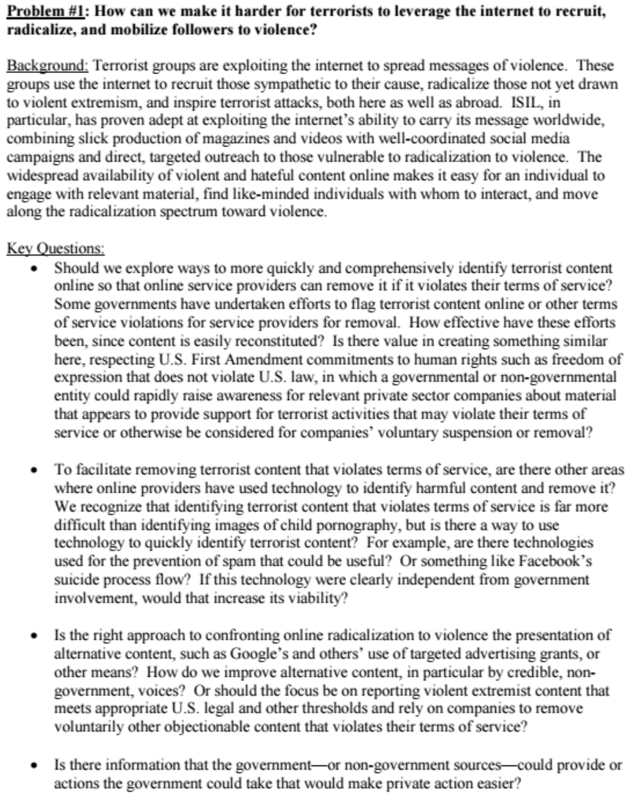
Algorithms that filter content are “a really powerful tool for a more authoritarian government,” said Schneier.

“Electronic monitoring and censorship can be effective for chilling political dissent, removing much content that authority frowns upon, and making people fearful of discussing political subjects online,” UC Davis’ Rogaway wrote in an email. “China already does this quite effectively.”

The government briefing does acknowledge that “respecting U.S. First Amendment commitments to human rights such as freedom of expression” would be important in any sort of system of identifying and reporting radical online posts.

But overall, the briefing document “reveals a troubling amount of magical thinking on the part of government officials,” German wrote. “It seems they are going to continue ignoring the vast amount of research that describes terrorism as a complex behavior that can only be understood in the context of the political situation in which it arises, and continue investing in snake-oil salesmen who promise a simple solution that identifies the ‘bad guys’ right before they strike.”

Read the rest of the document below:



Problem #1: How can we make it harder for terrorists to leverage the internet to recruit, radicalize, and mobilize followers to violence?

Background: Terrorist groups are exploiting the internet to spread messages of violence. These groups use the internet to recruit those sympathetic to their cause, radicalize those not yet drawn to violent extremism, and inspire terrorist attacks, both here as well as abroad. ISIL, in particular, has proven adept at exploiting the internet’s ability to carry its message worldwide, combining slick production of magazines and videos with well-coordinated social media campaigns and direct, targeted outreach to those vulnerable to radicalization to violence. The widespread availability of violent and hateful content online makes it easy for an individual to engage with relevant material, find like-minded individuals with whom to interact, and move along the radicalization spectrum toward violence.

Key Questions:

- Should we explore ways to more quickly and comprehensively identify terrorist content online so that online service providers can remove it if it violates their terms of service? Some governments have undertaken efforts to flag terrorist content online or other terms of service violations for service providers for removal. How effective have these efforts been, since content is easily reconstituted? Is there value in creating something similar here, respecting U.S. First Amendment commitments to human rights such as freedom of expression that does not violate U.S. law, in which a governmental or non-governmental entity could rapidly raise awareness for relevant private sector companies about material that appears to provide support for terrorist activities that may violate their terms of service or otherwise be considered for companies’ voluntary suspension or removal?

- To facilitate removing terrorist content that violates terms of service, are there other areas where online providers have used technology to identify harmful content and remove it? We recognize that identifying terrorist content that violates terms of service is far more difficult than identifying images of child pornography, but is there a way to use technology to quickly identify terrorist content? For example, are there technologies used for the prevention of spam that could be useful? Or something like Facebook’s suicide process flow? If this technology were clearly independent from government involvement, would that increase its viability?

- Is the right approach to confronting online radicalization to violence the presentation of alternative content, such as Google’s and others’ use of targeted advertising grants, or other means? How do we improve alternative content, in particular by credible, non-government, voices? Or should the focus be on reporting violent extremist content that meets appropriate U.S. legal and other thresholds and rely on companies to remove voluntarily other objectionable content that violates their terms of service?

- Is there information that the government—or non-government sources—could provide or actions the government could take that would make private action easier?