Artificial Intelligence and Counterterrorism: Possibilities and Limitations

House Homeland Security Committee, Subcommittee on Intelligence and Counterterrorism

Prepared Testimony and Statement for the Record of

Ben Buchanan

Assistant Teaching Professor, School of Foreign Service
Senior Faculty Fellow, Center for Security and Emerging Technology
Georgetown University

Thank you, Chairman Rose and Ranking Member Walker, for holding this important hearing and for inviting me to testify.

My name is Ben Buchanan. I am an Assistant Teaching Professor at the School of Foreign Service and a Senior Faculty Fellow at the Center for Security and Emerging Technology, both at Georgetown University. I am also a Global Fellow at the Woodrow Wilson International Center for Scholars, where I teach introductory classes on AI and cybersecurity for congressional staff. My research specialty is examining how cybersecurity and AI shape international security. --I co-authored a paper entitled "Machine Learning for Policymakers."[1]

The title of todays' hearing rightly alludes to the possibilities and limitations of AI as it applies to counterterrorism. To help structure our examination of both, I'd like to offer some thoughts to conceptualize the potential areas of contribution and concern.

AI as a Tool of Moderation

AI offers some promise as a tool of moderation. Social media platforms operate at a gigantic scale, sometimes including several billion users. It is deeply unrealistic to think that any team of humans will be able to monitor communications at that scale without automated tools. Social media moderation remains a thankless and grueling job for those individuals who do it, but AI is already useful in lessening the burden at least somewhat. Perhaps most importantly, AI sometimes helps platforms respond more quickly to objectionable content, swiftly preventing it from spreading. Optimists envision a platform in which AI quickly and effectively takes on the vast, or the entire, share of the difficult job of moderation, leaving users to enjoy an online experience that meets their expectations.

I am deeply skeptical that this is possible. In general, policymakers underestimate the power of machine learning systems and the rapid rate of change, but I think the moderation problem is one of the most fiendishly difficult ones--so difficult, in fact, that technology companies struggle to come up with enforceable and clear standards that their human moderators can consistently enforce, much less standards that machines can apply.

There are at least four reasons why this problem is hard. First is that context is vitally important, and context can often be hard for algorithms to grasp. The same video of a

[1] Buchanan, Ben and Taylor Miller. "Machine Learning for Policymakers." *Belfer Center for Science and International Affairs* (2017), https://www.belfercenter.org/sites/default/files/files/publication/MachineLearningforPolicymakers.pdf.

terrorist attack might be propaganda in one setting but legitimate news reporting in another. The same video of soldiers on patrol may in one format be meant to instill patriotic pride but in another context serve as a threat to those soldiers. My understanding of the technology is that it is a long way from being able to identify this context and respond appropriately.

Second is that machine learning based systems by definition rely on distilling patterns, and objectionable content does not always fit into neatly observable patterns. For example, content moderation systems are vastly less effective in unfamiliar languages. In addition, they are less effective at catching objectionable content that takes on unfamiliar forms. For example, one five-month study and whistleblower complaint obtained by the AP contends that Facebook, using both its automated and human moderation capabilities, removed only 38% of content posted by terrorist organizations. Facebook claims that its systems are much more effective, citing a 99% success rate, but it seems that the firm's denominator in calculating that percentage is only a subset of the content that is prohibited. The AP found that much objectionable content slipped through algorithmic filtering, including "an execution video, images of severed heads, propaganda honoring martyred militants."[2]

The third reason I am skeptical that AI can solve the moderation problem is that, sometimes, there is not sufficient time to train machine learning systems with new data. Consider the gun massacre in Christchurch, New Zealand, in which the objectionable content was streamed live all over the world. Such a thing had never been done before, and social media companies' automated systems were not nearly sufficient to keep the video from going viral, in part due to how users slightly edited the video to evade the detection of those systems. Unfortunately, we must assume that our adversaries will innovate and improve, finding weaknesses and exploiting them before companies have a chance to respond.

Fourth, and related, is that partial success with content moderation is often not sufficient. Consider the video of the terrible Christchurch shooting once more. As I said, though Facebook and YouTube were able to take down some copies, many other copies evaded their detection. The copies that did escape the filter were more than sufficient to ensure that the video still was able to go viral and attract widespread attention.[3]

---

[2] Butler, Desmond, and Barbara Ortutay, 'Facebook Auto-Generates Videos Celebrating Extremist Images', Wall Street Journal, 9 May 2019, https://www.apnews.com/f97c24dab4f34bd0b48b36f2988952a4

[3] Timberg, Craig, Drew Harwell, Hamza Shaban, and Andrew Ba Tran, 'The New Zealand Shooting Shows How YouTube and Facebook Spread Hate and Violent Images — Yet Again', Washington Post, 15 March 2019, https://www.washingtonpost.com/technology/2019/03/15/facebook-youtube-twitter-amplified-video-christchurch-mosque-shooting/?utm_term=.b37e9604a2da

In sum, to solve the moderation problem an AI system would have to not just identify content that might be objectionable, but also grasp context, be able to identify objectionable content that is distinct from what came before and in unfamiliar languages, respond quickly to an adversary's changing tactics, and work correctly a very large percentage of the time without a large number of false positives. I am skeptical whether such a system exists or will exist in the very near future. In short, from my vantage point, I see more limitations here than possibilities. I would encourage you to ask representatives of social media companies whether they think such a system is achievable or, if they do not think such a system is achievable, then what their plan is to scale content moderation to billions of users.

AI as a Tool of Radicalization

There is one other point to this discussion that I believe deserves significant attention: research has recently come out suggesting that automated recommendation systems can contribute to the radicalization of individuals. Such systems will recommend videos, articles, or other content to users on a social media platform based on what they have consumed on it already and what others users have viewed and liked, creating a loop of content designed to keep users on the platform.

Some academics argue that it is an effect by design for these recommendation systems, such as the Recommended Videos feature on YouTube, to push users towards even more extreme videos.[4] In this sense, then, AI is not a force for moderation online but in fact a force for radicalization.

My assessment of this research is that raises significant concerns, though I do not think the data is yet definitive. The research does, however, raise a very alarming possibility: that not only are automated moderation systems insufficient for removing objectionable content but that other automated systems--technology companies' recommendation algorithms--in fact are driving users to objectionable content that they otherwise would not find. If this is the case, then the technical limitations of AI work against the interests of national security, while its vast possibilities work for those who benefit from radicalization online. Again, the public data is limited and not yet conclusive, but it seems to me that the net effects of recommendation systems that steer users to content generated by others users need substantial additional study. I encourage you to ask technology companies about them.

---

[4] Nicas, Jack, 'How YouTube Drives People to the Internet's Darkest Corners', Wall Street Journal, 7 February 2018, https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478 Tufekci, Zeynep, 'YouTube, the Great Radicalizer', New York Times, 10 March 2018, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Worse still is that automated algorithms on social media platforms can not just drive users to objectionable content but help make that content more appealing and visible. The AP and academic researchers found that Facebook's algorithms automatically generate slick videos of some of the extremist content that has evaded its filtering. These algorithmically generated videos take images and videos that extremists have uploaded and package it to make it more neatly edited and synthesized--in essence, unintentionally doing the work of propaganda.[5]

Conclusion

In conclusion, we ought not to lose sight of a vital broader point: technology is not the reason we are here today. We are here because humans abuse a system that other humans have created. Human moderation as well as current and near-future technology are--in my view--insufficient to stop that abuse. My sense is that there is likely more that social media platforms could do to better manage the problem, such as reducing how quickly messages go viral, expanding their AI research teams focused on this issue, and adjusting their recommendation and generation algorithms, even if it comes at the expense of their business. That said, my best guess is that these steps might mitigate the problem but are unlikely to solve it.

I thank you again for holding this hearing and I look forward to your questions.

[5] Butler, Desmond, and Barbara Ortutay, 'Facebook Auto-Generates Videos Celebrating Extremist Images', Wall Street Journal, 9 May 2019, https://www.apnews.com/f97c24dab4f34bd0b48b36f2988952a4