Justsecurity.org

Disinformation, Radicalization, and Algorithmic Amplification: What Steps Can Congress Take?

by Ambassador (ret.) Karen Kornbluh
February 7, 2022

Editor's note: This article is part of a series from leading experts with practical solutions to democratic backsliding, polarization, and political violence.

Design features of social media platforms are exploited to promote extremism. The platforms' after the fact, whack-a-mole approach to content moderation is insufficient. However, Section 230 reform – a popular rallying cry – is a blunt instrument that may lead to unintended suppression of important speech and not address radicalization. Congress or agencies acting on their own should push for transparency, consumer protection (in terms of consistent, transparent enforcement of terms of service), and development of transparent codes of conduct. This approach mirrors elements of Europe's Digital Services Act and could be endorsed by the United States and European Union (EU).

A Glimpse Inside Online Amplification Engines

When an internal Facebook experiment created a fake account for a fictional user – "Carol Smith," a 41-year old conservative mother from North Carolina – this account was recommended pages and groups related to QAnon within days of its creation. "Carol" was recommended an account associated with the militia group Three Percenters within three weeks.

This research and other documents released by the Facebook whistleblower Frances Haugen underscore that the design features of large social media platforms are exploited to promote extremism.

An August 2019 Facebook internal memo admitted that, "the mechanics of our platform are not neutral" and, in fact, that core product mechanics, including virality, recommendations, and optimizing for engagement are key to why hate and misinformation flourish on the platform.

Research shows this is true as to other platforms as well. TikTok's recommendation algorithm likewise promotes content from QAnon, the Patriot Party, Oath Keepers, and Three Percenters. In one example, after users interacted with transphobic videos on TikTok, the recommendation algorithm fed these users videos with hate symbols, white supremacist and anti-Semitic content, as well as coded calls to violence.

The video sharing platform YouTube has 290 extremist channels, according to research by the Anti-Defamation League. When researchers experimentally suggested to the algorithm, through their viewing habits, that they were interested in militant movements, YouTube suggested videos to them with titles like "5 Steps to Organizing a Successful Militia" and "So You Want to Start a Militia?" The platform also recommended videos about weapons, ammunition, and tactical gear to what the researchers at the Tech Transparency Project call "the militia-curious viewer."

How does incendiary material, including misinformation, get boosted to viral status on social media platforms? Online outlets like Breitbart and Daily Wire cherry pick anecdotes that appear to validate conspiracy theories about vaccines or voter fraud. The material is shared by influencers like the conservative activist Candace Owens and networks of users (or bots) with their audiences of followers. Algorithms designed to keep users online then boost the material into feeds of users like "Carol" because it is popular – whether or not she chose to follow any of the accounts carrying it. Ads further target the material based on detailed profiles about users and the ads themselves are boosted if an algorithm determines they will keep people with a particular profile online so that, in the words of the Facebook whistleblower, "An ad that gets more engagement is a cheaper ad."

The more incendiary the material, the more it keeps users engaged, the more it is boosted by the algorithm. Haugen relayed that European politicians complained to Facebook that they were being forced to post more extreme messages online – and

therefore to adopt more extreme political positions – in order to get their messages heard under the platform's model of algorithmic amplification. When their messages were substantive, they did not reach as many users as when they attacked each other.

Other social media features were exploited for high-tech organizing of the Stop the Steal groups that culminated in the January 6th insurrection. "Super inviters" create invitation links – in part based on suggestions from the platform or pulled from lists of other groups – that can be shared on or off Facebook and can easily coordinate their invitations. A Facebook internal report on the Stop the Steal movement revealed that 0.3 percent of group members were responsible for 30 percent of invitations to join. Among these 137 "super-inviters" to the largest Stop the Steal group, Facebook reported that 73 were members of other groups designated as harmful conspiracy groups, which fed into high membership overlap with Proud Boy and militia groups and fueled Stop the Steal Groups' meteoric growth rates. Research shows Facebook even directs users who "like" one militia page toward other militia groups.

The research also makes clear that the companies' current strategy of post hoc, individual take-downs is grossly insufficient to address this systemic vulnerability. A COVID conspiracy video hosted by Breitbart's pages and channels on social media, America's Frontline Doctors, reached over 20 million viewers before all the major platforms took it down for violating their terms of service. Similarly, Stop the Steal, militia movements, and QAnon grew to reach millions of users before the mainstream platforms took concerted action to contain them.

Not Just a Facebook Problem: Extremist Amplification and its National Security Consequences

These design features pose a national security risk when foreign agents and terrorists use them to recruit and harass American citizens. The first volume of Special Counsel Robert Mueller's report documents Russians' "sweeping" and "systemic" social media disinformation in the 2016 U.S. presidential election. Russians used government-backed outlets, like RT and Sputnik, as well as fake and conspiratorial sites, and fake pages such as Blacktavist, to spread disinformation, amplifying stories with bots and trolls. They not only attempted to discourage African Americans from voting, inflame fears of immigrants, and spread disinformation about Hillary Clinton, but also organized groups

of Americans to take to the streets to protest each other, in one case in front of a mosque. More recently, during the 2020 election, Eastern European troll farms controlled Facebook's most popular pages for Christian and Black American content, reaching 140 million U.S. users a month. In September of 2021, a pro-China COVID-19 misinformation campaign spread over 30 social media platforms; analysts believe the primary objective was to physically mobilize protestors in the United States.

But the national security risk arises not only from foreign interference but also when domestic extremists use social media. In fact, the Federal Bureau of Investigation warns on its website, "[i]nternational and domestic violent extremists have developed an extensive presence on the Internet through messaging platforms and online images, videos, and publications. These facilitate the groups' ability to radicalize and recruit individuals who are receptive to extremist messaging." In fact, in the days surrounding the December 2020 election, domestic actors including elected officials actively promoted destabilizing misinformation through their social media channels, leading to threats of violence against elections workers and culminating in the events of January 6. Indeed, the January 6 insurrection was extensively planned by domestic actors on social media platforms – from Facebook and Twitter to more niche sites like "The Donald" and Parler – where extremist messages easily found new audiences, thanks to the engagement-based boosting of the platforms.

Platforms do not seem to be fixing the design defects so easily exploited by extremists. The events of January 6th, congressional hearings, and media exposés have yielded takedowns of QAnon, militia groups, and promoters of vaccine and election conspiracies. Facebook deemphasized political posts and current events in NewsFeed, recommitted to removing Taliban content, and reworked warning labels. Instagram will soon give the option for chronologically-ordered feeds. Twitter rolled out new features intended to make it easier for users to report Tweets. These are important steps, but they remain insufficient and piecemeal. Facebook's Oversight Board has been stymied in its review the company's "XCheck," or cross check, system that has exempted high-profile users from some or all of its rules due to a lack of transparency and unsatisfactory answers, according to the board's co-chair. Across the board, the platforms have not begun fundamental reform. Indeed, the

Facebook papers disclosed that the company rejected insightful employee ideas for changing design flaws to reduce the gravest harms from online extremism.

Platforms currently enjoy legal immunity for almost all content posted by or activity engaged in by others on their networks. They also face no requirement to be transparent about what is occurring on their platforms – except to report child pornography and child sexual exploitation. Until government changes social media companies' incentives, these sites will not change the design features that enable and promote algorithmic radicalization.

Commonsense Steps to Limit Algorithmic Radicalization

The recent debate about fixing social media has focused on Section 230 of the Communications Act of 1996, which shields social media platforms from liability for "content provided by another information content provider." Both Donald Trump, as president, and then-candidate Joe Biden threatened to cut back the law, and there are numerous bills in Congress to eliminate or change it.

In fact, carving back Section 230 protections for third party content is far from a panacea. Not only does the First Amendment robustly protect debate on issues of public concern, but even where a crime is alleged to have been committed online, in order for the platform itself to be liable, the underlying statute would also need to bar facilitation of the crime or distribution of the content.

As a result, removing Section 230 protection would allow cases to proceed only where platforms were alleged to have violated statutes that reach conduct of intermediaries (e.g., where the platform is alleged to have provided material aid to a designated international terrorist organization), or to have been negligent in allowing their sites to be used in ways that resulted in serious harm. Negligence could be alleged, for instance, in civil lawsuits against online service providers for failing to act against recruitment and planning for acts of violence that resulted in injury to plaintiffs (e.g., Capitol and DC Police Officers wounded on January 6th).

However, allowing one-off negligence lawsuits is a blunt instrument, opening up liability in a very broad range of cases that would take years to yield clear guidance for online

service providers about what constitutes appropriate safeguards. In the meantime, online service providers could well decide either to engage in overzealous moderation of nonviolent expressive content or conversely to wait until clarity arrives before implementing safeguards.

For this reason, any targeted Section 230 reform should be paired with positive transparency requirements, consumer protection investigations (e.g., of failure to enforce terms of service or of how violent extremist incidents were organized online), and development of a voluntary code of conduct.

Open the "Black Box"

Obtaining data on how social media sites both promote and moderate illegal and dangerous activity should not require the actions of a whistleblower. Congress and state legislatures can require specific types of data be provided to regulators, researchers, and the general public.

- Just as the National Transportation Safety Board gets access to data on airplane crashes or the Environmental Protection Agency releases data on pollution, regulators should have access to data related to a violent event that may have been organized online, in a way that respects First Amendment concerns.
- Likewise, large platforms should be required to provide information on their enforcement of their terms of service and permit third-party audits of this enforcement. This transparency should extend to their users as well: platforms should consistently inform users of the reasons behind any account or content moderation decision, and provide them a right to appeal. None of this need require disclosure of trade secrets or the details of editorial decisions. Even aggregate statistics would be illuminating. (The bipartisan Platform Accountability and Consumer Transparency Act (PACT) sponsored by Senators Thune (R-SD) and Schatz (D-HI) for example, would require online platforms to provide biannual reports of disaggregated statistics on content that has been removed, demonetized, or deprioritized.)
- Researchers should have access to retrospective data on online activity in a way that removes user identities to protect user privacy. The bipartisan Platform Accountability and Transparency Act bill, sponsored by Senators Coons (D-DE),

Portman (R-OH), and Klobuchar (D-MN), would create a process for the National Science Foundation to approve data requests from independent researchers. Social media companies would be required to provide data for approved proposals, subject to enforcement by the Federal Trade Commission (FTC). The FTC would establish appropriate privacy and cybersecurity safeguards.

Advertising Transparency and Know Your Customer

The bipartisan Honest Ads Act would provide the same transparency about ads that is required on broadcast for election ads. It should be supplemented by Know Your Customer rules that prevent dark money or foreign actor ad funding. Platforms would be required to have robust systems for archiving advertisements and information on how those ads were targeted that are searchable and sortable through an Application Programming Interface (API). This requirement could apply to all advertisements, as Phillip Howard, Laura Edelson, and others have proposed. However, it is important to be mindful that much of online organizing of violent acts does not use paid advertising.

Regulatory Investigations

The FTC or a group of State Attorneys General could use their existing consumer protection authorities – which protect against deceptive or unfair trade practices – in order to investigate platforms' potential material failures to honor commitments in their own terms of service or unfair actions in failing to guard against significant harms to consumers. (Potential examples of such deceptive or unfair practices, revealed by the Facebook papers, include exempting 5.8 million VIP users from content moderation standards, the abysmal rates at which Facebook removes hate speech — less than 5 percent of all hate speech on the platform is taken down — and systematically allowing content that violates terms of use to go viral before taking it down.) In the event that a case yields a settlement, the consent decree or other settlement agreement could require the defendant platform to take steps to mitigate risks of further consumer deception, report compliance, and even require third party audits or assessments of a platform's harm mitigation program in light of evolving risks.

Indeed, one such regulatory investigation may already be underway. According to the Wall Street Journal, Facebook revealed in October that it is "subject to government investigations and requests relating to a former employee's allegations and release of

internal company documents concerning, among other things, our algorithms, advertising and user metrics, and content enforcement practices, as well as misinformation and other undesirable activity on our platform, and user well-being."

Transparent Best Practices

In addition, the National Institute for Standards and Technology (NIST) or the FTC could work with platforms to develop a platform code of conduct or best practice guidance for mitigating design features that amplify illegal content or activity. (The PACT Act contains a provision to require NIST to develop a voluntary code.)

Such a framework would work best with legislation to create baseline accountability and transparency. However, in the likely event that Congress is unable to pass legislation, the FTC or State Attorneys General have predicates to pursue consumer protection investigations against platforms. And NIST or the FTC could consult with platforms to develop a voluntary, transparent code of conduct for amplification and content moderation processes.

There is no silver bullet to address algorithmic radicalization, but this problem is clearly serious enough to warrant a series of steps that would make a difference at the margins, while continuing to respect First Amendment values and requirements.

Image: WASHINGTON, DC – DECEMBER 01: Former Facebook employee Frances Haugen (L) listens during a hearing before the Communications and Technology Subcommittee of House Energy and Commerce Committee December 1, 2021 on Capitol Hill in Washington, DC. The subcommittee held a hearing on "Holding Big Tech Accountable: Targeted Reforms to Tech's Legal Immunity." (Photo by Alex Wong/Getty Images)

About the Author(s)

Ambassador (ret.) Karen Kornbluh

Amb. (ret.) Karen Kornbluh (@KarenKornbluh) is a Senior Fellow and Director of the Digital Innovation and Democracy Initiative at the German Marshall Fund of the United States and former U.S. Ambassador to the Organization for Economic Cooperation and Development.