SAMPLE MEMORANDUM

To: *Congressman Bennie Thompson* – House Select Committee to Investigate the January 6th Attack on the United States Capitol
From: *Joseph C. Hansen* – Research Assistant, Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute of International Studies *at* Monterey

Re: Combating Algorithm-Driven Radicalization on Social Media to Counter Domestic Extremists

The proliferation of extremist content on social media has been inadvertently augmented by algorithms designed to enable virality and promote user engagement. Without in-platform distinction between political/extremist content and normal content dividing virality/promotion potential, extremist groups will continue to be able to use social media to recruit, radicalize, and spread their message at scale.

Platforms Incentivize Engaging Content

Social media platforms' algorithms base everything from search results to what goes viral on user input. Far from when platforms like Facebook only showed you content your friends had posted previously, almost all platforms have shifted to a suggestions-based model where your unique user data is used to provide the content you are most likely to engage with. This data is increasingly specific and targets individuals on every aspect of their lives. Today's algorithms can guess a user's age, gender, location, habits, and political party. This information is not just crucial for advertisers to run highly targeted ad campaigns; it also helps the platforms keep users engaged by suggesting increasingly specific content related to previously demonstrated interests.

This dynamic means that platforms are not just incentivized to collect the data for advertisers; they are also inclined to use the same data to keep individual users on-platform for as much time as possible (to show them as many ads as possible). The culmination of the problem stems from the combination of new suggested content with previously available reposting features. This component allows a piece of content to both organically go viral and enables platforms to augment its virality by pushing it in suggested spaces in the app. It also incentivizes users to create viral content through community engagement and personal monetary benefits (both from platforms and 3rd parties looking for "organic" marketing tools).

While this structure works well for non-political content, values-related content is subject to the same dynamics. User engagement is the only important metric, independent of whether public reaction to the posted content is positive or negative. Hence, as engagement with relatively extreme (albeit within generally acceptable political discourse) content is higher, extreme political views are often augmented by platform's incentive structure.

Content Moderation Will Never Effectively Prevent the Spread of Extremist Content

In 2018, after several terrorist attacks in Europe, the United Kingdom sought to enforce content moderation in its "Counter-Terrorism Strategy." To stem the tide of extremist-adjacent influencers, they tried to enforce community standards against terrorist propaganda and radicalization. The strategy sought to hold Communication Services Providers (ISPs, in effect) responsible for allowing the transmittance of malfeasant content.

In the United States, great effort has also been made to track, report, and remove foreign extremist content, along with significant expansion of community standards to include a wider variety of content. While Section 230 provides shelter from liability, platforms do not want terrorist-adjacent content to remain as it would disincentivize engagement. *The problem is the focus on content moderation*, mostly from users flagging individual posts or accounts before AI then human oversight. While platforms have become more sophisticated with device and IP tracking, a banned user can still create another account and continue posting.

Even with AI assistance, the sheer quantity of content published daily is impossible to moderate effectively, especially given the increased usage of codewords and slang to bypass AI screeners. It will never be enough to simply moderate extremism off the internet; instead, *we must reduce its ability to go viral.*

Limit the Virality Potential of Extremist Content by Separating Algorithms

In place of moderation, algorithmic structure could be changed. Values-related content should not continue to be subject to the same virality dynamic and user incentive structure. This filtration could be achieved through text/audio screening AI looking for keywords generated by a platform's moderation team. Then, viral values-related content could be subject to limited virality, barring third-party verification of the veracity and authenticity of the content. A principal caveat to this system would be that politicians, national campaigns, and other verified users' content would be excluded from this screening process, reducing potential limiting effects on political speech.

The essential idea is that while community-generated extremist content would be stopped at the community level, it would not interfere with an individual's right to post it, nor their standing within their community/group. Instead, content that foments extremism (of whatever motivation, and particularly violence) is limited. In this system, said content:

- Cannot be shared/reposted more than five layers beyond the original creator.
- Duplicate content cannot be reposted from multiple accounts within the same group/by individual user(s).
- The algorithm will not separately promote this content in a suggested position in-app.
- Ads/Ad Data cannot be purchased for such content.

The secondary virality tier for political content could be bypassed via several means. The most important is user verification: that the user is indeed who they claim to be, that there is a single account associated with them (allowing more for businesses or anonymous authors, etc.), and that any qualifications they may have to produce authentic values-related content are established. This would also help counter foreign influences posing as members of domestic communities.

Mitigating Free-Speech Concerns: Federal Oversight & Public Confidence

Free-speech protections are at the center of this issue and in the past 5 years especially; right-wing groups have taken to attacking platforms for restrictive moderation policies. While mitigating the proliferation of extremist content is important, there are two principal concerns related to speech issues: potential inhibition of political movements and trust in private-sector moderators.

First, this broad-spectrum policy, applying to nearly all values-related content, will profoundly affect the ability of grassroots political movements to start. Mitigation can include verifying community-based organizations or increasing the scope of individuals not subject to the secondary filtering system. Though ultimately, limitation is part of the intent. To slow extremist movements from abusing social media virality, legitimate political movements will have to work harder to bypass initial screening before most legitimate content is no longer flagged. If designed with the right balance, strong and legitimate movements will not have significant difficulty bypassing the second tier and gaining traction. In contrast, extremist organizations will have their message significantly slowed in proliferating.

The second issue arising from implementation is increased intervention by social media companies in regulating political discourse. This will *require federal guidance* and protection. While the First Amendment currently protects social media companies' right to limit speech on their platforms (as they are considered private forums), this protection may not be unlimited,[1] and recently, platforms have become political targets. Low public confidence in these companies could drive the creation of new ones outside the bounds of this system. Transparency and public trust are going to remain incredibly important if implementation should take place.

[1] Person, & Trotta, D. (2022, September 17). *U.S. appeals court rejects Big Tech's right to regulate online speech*. Reuters. Retrieved October 10, 2022, from https://www.reuters.com/legal/us-appeals-court-rules-against-big-techs-ability-regulate-online-speech-2022-09-16/

Note: Additional Sourcing Information removed due to NDA, may be supplied on request.