The Double-Edged Sword of Banning Extremists from Social Media

Sam Jackson

University at Albany
sdjackson@albany.edu

Over the past few years, researchers, activists, and policymakers have engaged in debates over how social media companies should respond to extremism on their platforms. One facet of this debate focuses on the consequences – online and offline – of different approaches. Debates about the effectiveness of various approaches have not recognized that there are two different goals: reducing extremist violence and reducing extremism. This article presents a thought experiment that unpacks these goals, thinks through possible relationships between different approaches and different goals, and suggests a number of hypotheses that could be tested to empirically investigate the consequences of banning or tolerating extremists on social media.

In March 2018, Jack Dorsey, CEO of Twitter, announced a new initiative: Twitter would solicit proposals from experts on “conversational health” to try to “increase the collective health, openness, and civility of public conversation” online (jack, 2018a). Like other times that Jack has talked about problems of violence, intimidation, and extremism on his platform (for example, jack, 2018b, 2017) a number of Twitter users had a simple suggestion: just get rid of the Nazis.

Particularly since the rise of the so-called “alt-right,” there has been an increased conversation in America and around the world about how social media companies should respond to extremism on their platforms. Though some people identify seemingly simple solutions (like getting rid of all the Nazis), there is substantial disagreement about what action platforms should take. Debates have centered around three issues: (1) tension between protecting civil liberties and protecting public safety (Corynne McSherry in Abumrad, 2017; Cope et al., 2017); (2) the technical challenges of identifying extremists, and the lack of a transparency and an appeals process for those labeled as extremists (Alexander and Braniff, 2018; Baulke, 2016; Feamster, 2018; Ganesh, 2018; Knight, 2018; Moser, 2017); and (3) the consequences – on-platform and off-platform, online and offline – of banning extremism versus tolerating it.

This article takes up the third issue; given space limitations, I do not take up arguments about freedom of speech or other civil liberties, and I also do not take up arguments about technical difficulties and problems with processes of moderation. Instead, in the following pages, I argue that some of the disagreement over the consequences of responding to extremism online follows from the fact those who advocate for action might have two different goals – reducing

extremist violence and reducing extremism more broadly – but these goals are rarely articulated; a given response to extremism may have different consequences for these different goals. Making these goals explicit can add a measure of clarity to this debate. This article presents a thought experiment that unpacks a dramatically simplified version of this debate in an attempt to explore its complexities; as I will argue, untangling the consequences of platform moderation of extremism is incredibly complicated even given the simplifications assumed here, and attempting to unpack these complexities without these simplifications would render the following argument interminably long or simply incoherent.

After articulating the goals, I think through the disparate consequences of different responses on each goal. In closing, I suggest some possible strategies to empirically test the relationships that I propose here. Along the way, I identify a number of testable hypotheses, listed in the appendix. Indeed, this article’s primary contributions are unpacking a small portion of the complexity around this debate and developing hypotheses that could form the basis for future empirical work. Before any of this, though, I preface my argument with a brief discussion of what extremism is.

Defining Extremism

Despite increased attention to the problem of extremism, few scholars or commentators define extremism, trusting instead that their audiences will just know who is an extremist and who isn’t – as a recent book that provides an introduction to the concept of “extremism” observes, definitions of the term often follow common definitions of pornography: “we know it when we see it” (Berger, 2018a, p. 1). For those looking to better understand extremism and design policies to reduce it, though, that approach is not appropriate.

A few scholars have provided definitions of extremism. Some definitions focus on the use of violence (Breton et al., 2002; Midlarsky, 2011; Wintrobe, 2006); others portray extremism as any political ideology or movement that opposes core democratic principles (Backes, 2007; Mudde, 2014). J.M. Berger has recently argued that the defining feature of extremism is a belief that “hostile action against an out-group” is necessary for a group’s “success or survival” (Berger, 2018a, p. 170). Instead, I define political extremism as any attempt to change a fundamental feature of a political system [citation removed for blind review]; this definition is intentionally broad and is meant to encourage comparisons between a wide range of actors and behavior.

For the purposes of this article, though, settling on a single definition of extremism is not necessary. Despite the language used in public conversation over the issue of platform responses to extremism – in fact, despite the language used here – platforms do not respond to “extremism” no matter how that term is defined: they respond to particular forms of behavior and to particular users. This article discusses platform responses to extremism rather than particular behavior and users because the relationships between goals and approaches laid out here are content-neutral: they are likely as true for Salafi jihadist extremists as for male supremacy extremists, though the exact details of the effects on an extremist group of certain responses by platforms will change depending on the patterns of that group’s online activity (i.e., what platforms the group is active on, what purposes the group uses these platforms for, etc.).

Goals

This article is focused on the consequences of different responses by platforms to extremism online. Broadly speaking, there are two dominant goals for actors responding to extremism. The first goal is reducing extremist violence; the second is reducing extremism more broadly. These two goals parallel the distinction between disengagement (which is largely focused on leading individuals to stop participating in extremist activity, especially violence) and deradicalization (which is largely focused on leading individuals to renounce extremist beliefs) (Horgan, 2008). Disengagement is often a more modest goal compared to the broader goal of deradicalization. Similarly, reducing extremist violence is much more concrete than is reducing extremism.

Even the idea of reducing extremism is not all that clear. Does it mean weakening individuals’ extremist commitments? Shrinking extremist groups? Marginalizing extremist ideologies? In a sense, it is all of these. Marginalizing extremist ideologies might make communities that are identified with those ideologies themselves seem less acceptable, which likely would discourage some individuals from joining such groups (but would likely encourage others to join).¹ Shrinking extremist groups likely makes extremist ideas less prevalent: individuals seeking out that form of extremism would have a harder time finding the content they want, and

¹ For example, Cynthia Miller-Idriss’s study of German far-right youth (Miller-Idriss, 2017) reveals that some young people are drawn to right-wing extremism precisely because it is counter-cultural, an act of rebellion against the society they live in.

fewer individuals would be exposed to such content inadvertently.² Weakening engaged individuals' commitments to extremism might make these individuals less likely to spread their extremist ideas and recruit new members; it might also make them more receptive to deradicalization programs. For the purposes of this paper, "reducing extremism writ large" means reducing the prevalence of extremist ideas and the number of individuals affiliated with extremist groups and movements.

Reducing extremist violence is much more concrete than reducing extremism, but even it contains a number of different possibilities. It could mean reducing the number of violent crimes committed by extremists; or it could mean reducing the size and frequency of large rallies and marches that have an aggressive demeanor; or it could mean reducing harassment and intimidation online. In addition, it could happen through a variety of different pathways: convincing those who are part of extremist groups or who hold extremist ideologies that violence is not a legitimate way to pursue their political goals; making it more difficult for extremists who are in favor of violence to carry out attacks; or convincing those who are part of extremist groups or hold extremist ideologies that they are not so threatened that violence is warranted at this point. As Busher, Holbrook, and Macklin observe, extremist groups might even decide themselves to avoid violence (Busher et al., 2019).

Of course, these two primary goals are not inherently contradictory: for example, reducing extremism writ large might (or might not) reduce the amount of extremist violence. But, as I will argue, some approaches used to pursue one goal may impede the pursuit of the other goal. In the next section, I further set the stage for that argument by describing approaches that may be used in response to extremism online.

Approaches

There are a variety of different approaches that platforms can use in response to extremism. Some approaches are aimed at the consumer of extremist content: for example, the Redirect Method uses online advertising technology to redirect individuals searching for ISIS

² One of the most prominent examples of an individual inadvertently coming across extremist material is Dylann Roof. In response to media reports about Martin Zimmerman's shooting of Trayvon Martin in 2012, Roof turned to the internet to learn more about the incident. After reading about the shooting, he searched for the phrase "black on white crime." He came across the Council of Conservative Citizens, a modern-day incarnation of the white supremacist White Citizens Councils that organized in defense of Jim Crow laws and in opposition to the Civil Rights Movement in the 1960s and 70s (Hersher, 2017).

content to counter-ISIS content instead (“The Redirect Method,” n.d.). Other approaches are aimed at producers and sharers of content. This paper focuses on those approaches that fall on a spectrum of removing content along with those who produce or share it, from removing extremists entirely to removing specific extremist content to tolerating extremist content and users entirely (see figure 1). Platforms have experimented with intermediate approaches, including Twitter’s efforts to deprioritize certain results in searches and YouTube’s ongoing efforts to “demonetize” videos associated with extremist content (Harvey and Gasca, 2018; Walker, 2017). For the sake of simplicity, this paper focuses on the two ends of the spectrum: banning (some) extremist users and tolerating extremist users and content. A premise of this article is that there is insufficient clarity in the debate about how to respond to extremism; my goal is to increase clarity through a simplified version of the debate. Future work will examine the debate in greater complexity.

Figure 1

How to Reduce Extremist Violence

Next, I consider the relationship between these goals and approaches, starting with the goal of reducing extremist violence. With the approach of tolerating extremists, it is possible that these individuals will be exposed to differing opinions, keeping them out of an echo chamber of awful. This could lead the extremism espoused by an individual to seem exceptionally extreme in comparison with other non-extremist ideas and activity advocated by other users on that platform. For example, if a user were to advocate for some sort of relationship between biology and national culture on Twitter (perhaps arguing that someone’s skin color or nose shape means that they cannot be a “real American”), that idea seems relatively extreme in comparison with

more common depictions of visually-diverse American scenes. In another example, if an individual were to advocate using violence to pursue her political goals in the streets of Charlottesville, that behavior seems relatively extreme in comparison with venerated examples of non-violent activism that carry the most legitimacy in contemporary America. In this context, perhaps a torchlit march seems extreme enough, with more direct violence seeming unnecessary or inappropriate.

With the approach of removing extremists, the dynamic could be different. Instead of placing extremist content and users in more diverse settings with moderate and even opposing voices, removing extremists from mainstream platforms could push them into more marginal online spaces (like Gab and 4chan) (Ellis, 2016; Marwick and Lewis, 2017; Rao, 2016; Zannettou et al., 2018). In these marginal spaces, they would have less interaction with moderate and opposing content, instead finding themselves in an echo chamber of awful where they might compete with each other to suggest a “truly” extreme idea or behavior. Without the moderating influence of advocates of non-violence, perhaps violence would seem like a more reasonable way to pursue one’s political goals. Inhabiting this homogenous space may also lead extremists to think that their political values are common among the public at large (Pitcavage, 2018), imagining themselves as the vocal vanguard of a silent majority. On top of this difference in context, being pushed off mainstream platforms and into marginal spaces likely increases the sense of persecution by extremists, fueling their perception of grievances and their desire to address those grievances by any means necessary (Hwang, 2017).

How to Reduce Extremism

Next, consider the goal of reducing (a particular form of) extremism more broadly. With the approach of tolerating extremists, being in a context with non-extremists may provide extremist users with more opportunities to recruit new supporters. Just as importantly, it may provide them with more opportunities to inject their ideas into broader political conversations, allowing them to shift the Overton window (the range of ideas and rhetoric deemed acceptable in a political community) (Russell, 2006). At the same time, allowing these extremist users to remain on mainstream platforms may provide more opportunities for deradicalization or disengagement through interaction with opposing ideas. However, deradicalization and disengagement via interaction on social media at scale is an unlikely prospect: it is hard enough to identify

extremists, much less individuals who might be susceptible to extremism but are not necessarily part of extremist groups or movements yet; and it is harder still to design interventions that actually accomplish their goals (an embarrassing example of an ineffective intervention is the State Department’s “Think Again Turn Away” Twitter campaign, in which a Twitter account with clear U.S. State Department branding would tweet at ISIS-related accounts) (Miller and Higham, 2015). Put simply, whether tolerating extremists can reduce extremism hinges on whether extremist propaganda or counter-extremist propaganda is more effective.

The approach of removing extremists might see the opposite dynamic on both of these dimensions. That is, removing extremists might provide them with fewer opportunities to recruit new supporters or change political conversations (Berger, 2016a; Berger and Morgan, 2015; Berger and Strathearn, 2013). It likely provides them with fewer opportunities to spread misinformation, engage in trolling, or harass their opponents (Berger, 2016b; Herrman, 2016). In response to attempts by major social media platforms to develop automated tools to identify and remove extremist content, some extremist users have adopted the strategy of posting links to extremist content hosted on other platforms rather than posting the extremist content itself (Conway et al., 2017); removing those users would further disrupt these communication networks (Macdonald, 2018). It also might provide fewer opportunities for deradicalization or disengagement through interactions with non-extremist users; but as such interventions are difficult and are in their infancy, it seems unlikely that these fewer opportunities will translate into substantially fewer cases of actual deradicalization or disengagement.

The Relationship between Goals and Approaches

Given this rudimentary exploration of the possible consequences of different approaches to the two goals of reducing extremist violence and reducing extremism more broadly, we can fill in a preliminary typology (see table 1). In short, it seems plausible that tolerating extremist users could reduce extremist violence, and it seems less likely that banning extremist users would reduce violence. It seems unlikely that tolerating extremist users would reduce extremism, and it seems likely that removing extremist users would reduce extremism (in particular, given the likelihood that this could disrupt propaganda and recruitment networks). Ultimately, these are hypothetical answers to empirical questions, though.

		Goal	
		Reduce Extremist Violence	Reduce Extremism
Approach	Tolerate extremist users	plausible	unlikely
	Remove extremist users	unlikely	likely

Table 1: A Typology of Goals and Approaches

Complicating the Idea of Reducing Extremist Violence

In the previous discussion, I assumed that the reducing extremist violence happens at the collective level: that is, banning (or tolerating) extremist users affects levels of extremist violence because of the effects of that action at the group level. But in fact, it seems likely that actions I have described here work at the individual level. Overall levels of extremist violence are better understood as a combination of (among other things) individual propensity for violence and the number of individual extremists who have different propensities for violence.

Previously, I suggested that banning extremists might make violence more likely while tolerating extremists might make violence less likely. It might be more helpful to think about this in a slightly different manner: banning an extremist might make that individual more likely to use violence, while tolerating an extremist might make that individual less likely to use violence.

But, if we assume that tolerating extremist users allows extremists to do more recruiting (and recruiting, of course, increases the total number of extremists, if the number of new recruits exceeds those who leave extremism), then it may be that tolerating extremists decreases each individual's propensity for violence but increases the overall likelihood of violence.

To illustrate this, consider two hypothetical cases. In the first, social media platforms ban members of an extremist organization. In this case, the extremists who have been banned score a 5 on some hypothetical measure of propensity for violence. These extremists are part of a small group – say, 20 members. If all those members have the same propensity for violence, that suggests an overall propensity for violence score of 100.

In the second scenario, social media platforms do not ban these extremists. In this case, the extremists who have not been banned receive a 1 on that same hypothetical measure of propensity for violence. But, since these extremists remain on major platforms and are able to

continue recruiting, they have a larger group – say, 200 members. In this case, if all these members have the same propensity for violence, that suggests an overall score of 200.

We can translate this into an equation:

$$v = v_i \times i$$

where

v is the overall likelihood of violence at a given time,

v_i is the average individual propensity for violence at a given time, and

i is the number of individuals at a given time.

In the case of banning extremists, then, v_i may increase while i may decrease, in which case v is determined by the relative increase of v_i and decrease of i . It might be that banning extremists decreases the number of extremists to such an extent that it offsets the increase in individual propensity for violence caused by the banning. But that is an empirical question.

Additional complications

As I have noted throughout this article, I have dramatically simplified the issue of responding to extremism online. I have already pointed to some of these simplifications: a focus on one facet of the debate around this issue; a focus on responses that target producers and sharers but not consumers of extremist content – and even then, a focus on two opposite ends of the spectrum of responses that remove extremist users or extremist content.

My focus on producers rather than consumers of extremist content points to another complication. This argument discusses extremist users who are the targets of action or non-action. But any response to extremism by a platform is likely to target a subset of the extremists on that platform: some extremist users will see their allies be suspended or banned (if a platform takes that approach) but will not be suspended or banned themselves. How will that action taken against their friends affect these individuals who remain on the platforms?³

My argument has also not examined how platforms' responses to extremism might affect the wider public, some of whom may be susceptible to extremist recruitment. If Facebook chooses to ban all so-called “alt-right” activists, will more members of the public conclude that

³ My thanks to Elizabeth Pearson for raising this question.

Facebook is violating the “alt-right’s” free speech or free assembly rights? Will they become convinced that Facebook is biased against a set of legitimate political views? Will they simply decide to go out and learn more about these individuals that Facebook has banned? If this action by Facebook has any of these consequences, that could lead to increased numbers of extremists (though new research suggests that the short-term bump in attention that results from being banned from major platforms is minor relative to the losses associated with being banned) (Koebler, 2018).

My argument here has also not considered how one platform’s response to extremism might affect extremists on another major platform. I suggested that banning extremists from mainstream platforms might push extremists to marginal spaces like Gab; but what happens if Facebook bans a particular user or group but Twitter does not? This complicates some of the consequences for each approach that I have laid out here. For example, if Facebook bans a particular user, that user can no longer use Facebook for recruitment and spreading propaganda; but if Twitter does not ban that user, does the user simply shift some of the recruiting and propaganda activity that previously happened on Facebook over to Twitter? It seems likely that the affordances of different platforms could limit the ability of users to move activity from one platform to another (Berger, 2018b, p. 45), but some platforms are similar enough that extremists might be able to move from one to the other with relatively small effects on their behavior, particularly if they already have existing networks of followers and supporters on other platforms.

At this stage, I do not have even hypothetical answers to any of these additional complications. More careful thought is needed to tease out the implications of all these different facets of the problem of extremism online.

Moving the debate forward

Given the argument I make in here – in particular, the testable predictions I make (see the appendix) – the next step would be to empirically investigate the effects that different responses to extremism have. At this point, designing such investigation is a tall task. Platforms are scrambling to decide how to respond to extremism (in part, because of pressure by governments, users, and advertisers) (“Germany to enforce hate speech law,” 2018; Maheshwari, 2018), and many of their approaches seem to be based on trial and error. They often focus on relatively clear cut cases, for example where countries have legally prohibited certain extremist

groups (like ISIS and al-Qaeda). Data on this subject are hard to come by, and the approaches adopted by some platforms hamper retrospective attempts to gather data (for example, when Twitter deletes a user and purges all of that user's data; if a researcher was not already collecting the available data about that user, it would be impossible to reconstruct that data).

In particular, testing the model of the likelihood of extremist violence as I have described it here might be impossible. Counting the number of individuals affiliated with extremism is hard enough – accurate counts simply do not exist for most forms of extremism. Measuring an individual's propensity for violence would be challenging at best (Berger, 2013). Even if such measurement was possible, generating reliable measures at scale would prove enormously difficult and costly.

Conclusion

The complexity of the dynamics I have described here are substantial, but they should not lead platforms to give up on interventions aimed at those who produce and share extremist content. Some observers continue to suggest that the best way to combat extremism online is with more non-extremist content, in a direct analogue to the argument that the best way to combat misinformation is with more true information. But I would argue that the past three years in particular have demonstrated that more speech is not the answer to problems of bad speech online, whether that bad speech is extremism or misinformation (Napoli, 2018).

Particularly in a context where political affiliation (whether that affiliation is with a political party, a social movement organization, or even a religious group with strong political positions) increasingly shapes how individuals respond to information, the nudges provided by counter-speech likely have limited effects, or they might even backfire (Bail et al., 2018).⁴

Similarly, the past three years in particular have shown that neutrality is not effective when responding to extremism. Platforms can push the responsibility for making decisions about which users and content should be removed to policymakers; in the current political context, that means that platforms will focus their moderation efforts on major terrorist organizations like ISIS and al-Shabaab, neglecting a host of domestic (and especially far right) extremist groups – groups that governments rarely criminalize but are still involved in violent political conflict and other

⁴ See also the extensive bodies of literature on selective exposure (for example, Chandler and Munday, 2016; Knobloch-Westerwick et al., 2015) and motivated reasoning (for example, Nyhan and Reifler, 2010).

problems associated with political extremism. Platforms have recently struggled to figure out how to take action against harmful content while minimizing accusations of political bias (for example, see responses by Twitter to accusations of being biased against conservatives in America) (Cellan-Jones, 2018; Stewart, 2018). But platforms have no legal responsibility to host content from all (legal) political perspectives – they have the right to remove content in accordance with their terms of service. I do not suggest that platforms should aim to become echo chambers, hosting only political content from a narrow range of perspectives. But I do believe that platforms should take clear and bold positions to disallow certain forms of legal but harmful content on their platforms. For example, Facebook could decide that qanon conspiracies (Bank et al., 2018) are harmful to society and will not be allowed on the platform.

Fundamentally, though, we should not want platforms to take action against extremists until they decide what their goals are. If the relationships between approaches and goals as described here are accurate, reasonable responses to extremism could simultaneously increase extremist violence while reducing extremism writ large (or vice versa). Platforms should decide what they want to accomplish; only then should they design large-scale interventions to achieve those goals. But these social media companies should not use this as an excuse to never take action (or to only take action in a haphazard way) – given their prominence and the role they place in contemporary public life, they have a responsibility to consider that role and how they can regulate their platforms to ensure that the role doesn't become overwhelmingly negative.

Platforms have begun to form partnerships to address the closely related problem of terrorism. For example, Tech Against Terrorism is an organization “supporting the tech industry tackle terrorist exploitation of the internet.”⁵ Similarly, the Global Internet Forum to Counter Terrorism (GIFCT) provides a mechanism for companies to share technical resources to identify and remove terrorist content.⁶ Given the cross-platform dynamics that are possible with moderation efforts aimed at extremist users and content, platforms should consider forming a network to share tools and collectively develop policies. Even if different platforms make different moderation decisions, such a network might help companies to anticipate potential problems or other developments resulting from the moderation activities of other platforms. Of course, not all social media platforms will be interested in such an effort: Gab's alleged commitment to a

⁵ <https://www.techagainstterrorism.org/>

⁶ <https://www.gifct.org/>

maximal understanding of free speech (and its more practical and obvious commitment to hosting far right content and users) likely rule out that company's participation. Still, an organized effort of whatever scope could allow platforms to make more thoughtful and coordinated decisions about how to respond to extremism.

This discussion has been premised in part on an assumption that platforms should want to take action against extremism. Certainly, though, social media companies might have financial incentives to avoid taking additional steps towards moderation. As mentioned before, platforms sometimes face political repercussions for moderation activities that could lead a substantial number of their users to leave the service. And, given that the size of the user base contributes to the market valuation of many social media companies, choosing to remove users could reduce the value of companies (Wang, 2017). If these companies sell themselves to investors as behemoths that facilitate connections and communication of all types, moderation leading to a reduced user base might impact their financial status; if companies instead choose to position themselves as large companies who take seriously their social and political roles and work hard to make sure their platforms aren't abused, perhaps there would be fewer financial repercussions for removing toxic users.

Despite these challenges, social media companies must take action in response to extremist users and content on their platforms. They should be aware of the complications inherent in moderation decisions, but they should not allow this to deter them. They should commit to developing careful policy that recognizes these complications; ideally, they should form networks to develop such policy collaboratively. This article has aimed to explore complicated ideas carefully; hopefully, platforms can use this exploration to develop thoughtful moderation policies.

Appendix: Testable Hypotheses

1. The relationship between goals and approaches does not depend on the content of the extremism that is being responded to. That is, the effects of moderation are content-neutral.
2. Marginalizing extremist ideologies makes those who adhere to those ideologies seem less acceptable.
 - a. Making extremist ideologies seem less acceptable discourages some individuals from joining related extremist groups.
 - b. Making extremist ideologies seem less acceptable encourages some other individuals who are motivated by a desire to be counter-cultural to join related extremist groups.
3. Shrinking extremist groups makes extremist ideas less prevalent.
 - a. Shrinking extremist groups makes it harder for individuals seeking out that form of extremism to find the content they want.
 - b. Shrinking extremist groups makes it less likely that individuals will be inadvertently exposed to that type of extremist content.
4. Weakening engaged individuals' commitments to extremism makes these individuals less likely to spread their extremist ideas.
 - a. Weakening engaged individuals' commitments to extremism makes these individuals less likely to recruit new supporters.
5. Weakening engaged individuals' commitments to extremism makes these individuals more receptive to deradicalization programs.
6. Reducing extremism writ large reduces the amount of extremist violence.
7. Tolerating extremists leads to the following consequences:
 - a. Extremists will be exposed to differing opinions.
 - b. A particular set of beliefs or actions will seem more extreme within the universe of diverse perspectives.
 - c. Extremists will have more opportunities to recruit new supporters.
 - d. Extremists will have more opportunities to inject their ideas into broader political conversations.
 - e. There will be more opportunities for extremists to experience deradicalization and disengagement programs.
8. Removing extremists leads to the following consequences:
 - a. Extremists will move to more marginal online spaces.
 - b. Extremists will have less interaction with moderate and opposing content.
 - c. Extremists will compete with each other to suggest more extreme ideas and actions.
 - d. Extremists will view violence as more reasonable.
 - e. Extremists will think their perspective is more common than it is.
 - f. Extremists will have an increased sense of persecution.
 - g. Extremists will have an increased desire to address their grievances by any means necessary.
 - h. Extremists will have fewer opportunities to recruit new supporters.
 - i. Extremists will have fewer opportunities to change political conversations.
 - j. Extremists will have fewer opportunities to spread misinformation.
 - k. Extremists will have fewer opportunities to harass their opponents.
 - l. Cross-platform communication networks will be disrupted.

- m. There will be fewer opportunities for deradicalization or disengagement.
- 9. Tolerating extremist users will reduce extremist violence.
- 10. Tolerating extremist users will not reduce extremism.
- 11. Banning extremist users will not reduce extremist violence.
- 12. Banning extremist users will reduce extremism.
- 13. Removing an individual extremist will increase that individual's propensity for violence.
- 14. The overall likelihood of extremist violence is a function of the number of extremists and each extremist's propensity for violence at a given time.
- 15. Removing extremist users leads to the following consequences for violence:
 - a. Increases those individuals' propensity for violence.
 - b. Decrease the total number of extremists through a reduction in recruitment.
 - c. Results in a lower overall likelihood of violence.
- 16. Removing specific extremist individuals affects other individual extremists who are not removed as well.
- 17. Removing extremists affects the wider public.

Bibliography

- Abumrad, J., 2017. The Hate Debate, More Perfect.
- Alexander, A., Braniff, W., 2018. Marginalizing Violent Extremism Online [WWW Document]. Lawfare. URL <https://www.lawfareblog.com/marginalizing-violent-extremism-online> (accessed 2.12.18).
- Backes, U., 2007. Meaning and Forms of Political Extremism in Past and Present. *Středo Evropské politické studie* 242–262.
- Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A., 2018. Exposure to opposing views on social media can increase political polarization. *PNAS* 201804840. <https://doi.org/10.1073/pnas.1804840115>
- Bank, J., Stack, L., Victor, D., 2018. What Is QAnon: Explaining the Internet Conspiracy Theory That Showed Up at a Trump Rally. *The New York Times*.
- Baulke, C., 2016. The Nature of the Platform: Dealing with Extremist Voices in the Digital Age. Mackenzie Institute. URL <http://mackenzieinstitute.com/nature-platform-dealing-extremist-voices-digital-age/> (accessed 2.12.18).
- Berger, J.M., 2018a. Extremism. The MIT Press, Cambridge, MA.
- Berger, J.M., 2018b. The Alt-Right Twitter Census. *VOX-Pol*.
- Berger, J.M., 2016a. Nazis vs. ISIS on Twitter: A Comparative Study of White Nationalist and ISIS Online Social Media Networks. George Washington University Program on Extremism.
- Berger, J.M., 2016b. The Social Apocalypse: A Forecast [WWW Document]. IntelWire. URL <http://news.intelwire.com/2016/08/the-social-apocalypse-forecast.html> (accessed 2.12.18).
- Berger, J.M., 2013. The Hate List. *Foreign Policy*.
- Berger, J.M., Morgan, J., 2015. The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter. Brookings Institution.
- Berger, J.M., Strathearn, B., 2013. Who Matters Online: Measuring influence, Evaluating Content and Countering Violent Extremism in Online Social Networks. ICSR.
- Breton, A., Galeotti, G., Salmon, P., Wintrobe, R. (Eds.), 2002. Political Extremism and Rationality. Cambridge University Press, New York.
- Busher, J., Holbrook, D., Macklin, G., 2019. The internal brakes on violent escalation: a typology. *Behavioral Sciences of Terrorism and Political Aggression* 11, 3–25. <https://doi.org/10.1080/19434472.2018.1551918>
- Cellan-Jones, R., 2018. Trump issues warning to internet giants. *BBC News*.
- Chandler, D., Munday, R., 2016. A dictionary of media and communication.
- Conway, M., Khawaja, M., Lakhani, S., Reffin, L., Robertson, A., Weir, D., 2017. Disrupting Daesh: Measuring Takedown of Online Terrorist Material and its Impacts. *VOX-Pol*.
- Cope, S., York, J.C., Gillula, J., 2017. Industry Efforts to Censor Pro-Terrorism Online Content Pose Risks to Free Speech [WWW Document]. Electronic Frontier Foundation. URL <https://www.eff.org/deeplinks/2017/07/industry-efforts-censor-pro-terrorism-online-content-pose-risks-free-speech> (accessed 2.18.18).
- Ellis, E.G., 2016. Gab, the Alt-Right's Very Own Twitter, Is The Ultimate Filter Bubble [WWW Document]. WIRED. URL <https://www.wired.com/2016/09/gab-alt-rights-twitter-ultimate-filter-bubble/> (accessed 2.13.18).

- Feamster, N., 2018. Artificial Intelligence and the Future of Online Content Moderation. VOX - Pol. URL <https://www.voxpol.eu/artificial-intelligence-and-the-future-of-online-content-moderation/> (accessed 9.14.18).
- Ganesh, B., 2018. Tech That Counters Online Islamic Extremism Must Also Focus On Right-Wing Extremism. Centre for Analysis of the Radical Right. URL <https://www.radicalrightanalysis.com/2018/08/22/tech-that-counters-online-islamic-extremism-must-also-focus-on-right-wing-extremism/> (accessed 9.14.18).
- Germany to enforce hate speech law, 2018. . BBC News.
- Harvey, D., Gasca, D., 2018. Serving healthy conversation [WWW Document]. Twitter. URL https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html (accessed 9.9.18).
- Herrman, J., 2016. Who's Responsible When Extremists Get a Platform? The New York Times.
- Hersher, R., 2017. What Happened When Dylann Roof Asked Google For Information About Race? [WWW Document]. NPR.org. URL <https://www.npr.org/sections/thetwo-way/2017/01/10/508363607/what-happened-when-dylann-roof-asked-google-for-information-about-race> (accessed 9.15.18).
- Horgan, J., 2008. Deradicalization or Disengagement?: A Process in Need of Clarity and a Counterterrorism Initiative in Need of Evaluation. *Perspectives on Terrorism* 2, 3–8.
- Hwang, J.C., 2017. Analysis | Why banning 'extremist groups' is dangerous for Indonesia. Washington Post.
- jack, 2018a. We're committing Twitter to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress. @jack. URL <https://twitter.com/jack/status/969234275420655616> (accessed 9.6.18).
- jack, 2018b. Fundamentally, we need to focus more on the conversational dynamics within Twitter. We haven't paid enough consistent attention here. Better organization, more context, helping to identify credibility, ease of use. Challenging work and would love to hear your thoughts and ideas. @jack. URL <https://twitter.com/jack/status/1020767835667120128> (accessed 9.6.18).
- jack, 2017. thread. We need to be a lot more transparent in our actions in order to build trust. @jack. URL <https://twitter.com/jack/status/918508443631108096> (accessed 9.6.18).
- Knight, W., 2018. Three problems with Facebook's plan to kill hate speech using AI [WWW Document]. MIT Technology Review. URL <https://www.technologyreview.com/s/610860/three-problems-with-facebooks-plan-to-kill-hate-speech-using-ai/> (accessed 9.14.18).
- Knobloch-Westernick, S., Johnson, B.K., Westernick, A., 2015. Confirmation Bias in Online Searches: Impacts of Selective Exposure Before an Election on Political Attitude Strength and Shifts. *J Comput-Mediat Comm* 20, 171–187. <https://doi.org/10.1111/jcc4.12105>
- Koebler, J., 2018. Social Media Bans Actually Work. Motherboard. URL https://motherboard.vice.com/en_us/article/bjbp9d/do-social-media-bans-work (accessed 9.9.18).
- Macdonald, S., 2018. How tech companies are successfully disrupting terrorist social media activity [WWW Document]. The Conversation. URL <http://theconversation.com/how-tech-companies-are-successfully-disrupting-terrorist-social-media-activity-98594> (accessed 9.13.18).

- Maheshwari, S., 2018. Revealed: The People Behind an Anti-Breitbart Twitter Account. The New York Times.
- Marwick, A., Lewis, R., 2017. Media Manipulation and Disinformation Online. Data & Society Research Institute.
- Midlarsky, M.I., 2011. Origins of Political Extremism: Mass Violence in the Twentieth Century and Beyond. Cambridge University Press, New York.
- Miller, G., Higham, S., 2015. In a propaganda war against ISIS, the U.S. tried to play by the enemy's rules [WWW Document]. Washington Post. URL https://www.washingtonpost.com/world/national-security/in-a-propaganda-war-us-tried-to-play-by-the-enemys-rules/2015/05/08/6eb6b732-e52f-11e4-81ea-0649268f729e_story.html (accessed 9.13.18).
- Miller-Idriss, C., 2017. The extreme gone mainstream: commercialization and far right youth culture in Germany. Princeton University Press.
- Moser, B., 2017. How Twitter's Alt-Right Purge Fell Short [WWW Document]. Rolling Stone. URL <https://www.rollingstone.com/politics/news/how-twitters-alt-right-purge-fell-short-w514444> (accessed 2.13.18).
- Mudde, C., 2014. Introduction: Political Extremism - Concepts, Theories and Democratic Responses, in: Mudde, C. (Ed.), Political Extremism. SAGE, Los Angeles, pp. xxiii–xxix.
- Napoli, P.M., 2018. What If More Speech Is No Longer the Solution: First Amendment Theory Meets Fake News and the Filter Bubble. Fed. Comm. L.J. 70, 55.
- Nyhan, B., Reifler, J., 2010. When Corrections Fail: The Persistence of Political Misperceptions. Political Behavior 32, 303–330. <https://doi.org/10.1007/s11109-010-9112-2>
- Pitcavage, M., 2018. 1. I had an interesting conversation with a reporter the other day and he mentioned in passing that a lot of extremist movements seemed to have dark or even apocalyptic visions of the future. I said that this was true and noted that extremists (of whatever type--left, right, pic.twitter.com/CziHFscZ0h. @egavactip. URL <https://twitter.com/egavactip/status/962823267789795329> (accessed 2.12.18).
- Rao, A., 2016. Social Media Companies Are Not Free Speech Platforms [WWW Document]. Motherboard. URL https://motherboard.vice.com/en_us/article/4xa5v9/social-media-companies-are-not-free-speech-platforms (accessed 2.13.18).
- Russell, N.J., 2006. An Introduction to the Overton Window of Political Possibilities [WWW Document]. URL <http://www.mackinac.org/7504> (accessed 4.29.18).
- Stewart, E., 2018. Trump buys into the conspiracy that social media censors conservatives [WWW Document]. Vox. URL <https://www.vox.com/policy-and-politics/2018/8/18/17749450/trump-twitter-bias-alex-jones-infowars> (accessed 9.15.18).
- The Redirect Method [WWW Document], n.d. . The Redirect Method. URL <http://redirectmethod.org> (accessed 9.13.18).
- Walker, K., 2017. Four steps we're taking today to fight terrorism online [WWW Document]. Google. URL <https://www.blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/> (accessed 9.9.18).
- Wang, S., 2017. Twitter Is Crawling With Bots and Lacks Incentive to Expel Them. Bloomberg Businessweek.
- Wintrobe, R., 2006. Rational Extremism: the Political Economy of Radicalism. Cambridge University Press, New York.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringhini, G., Blackburn, J., 2018. What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber? Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18 1007–1014. <https://doi.org/10.1145/3184558.3191531>