The Censorship Effect

An analysis of the consequences of social media censorship and a proposal for an alternative moderation model

v1.0 | 3/10/2022

minds.com/change

Authors

Bill Ottman, Co-founder, CEO, Minds

Daryl Davis, Race Reconciliator

Jack Ottman, Co-founder, COO, Minds

Jesse Morton, Founder, Parallel Networks

Justin E. Lane, D.Phil, CEO, CulturePulse Inc.

Prof. F. LeRon Shults, Ph.D., Ph.D., University of Agder; CRO, CulturePulse Inc.

Editors

Dr. Sophia Moskalenko

James Daly

Julian Rapaport, Ph.D.

Mark Ryan Sallee

Naama Kates

Dr. Nafees Hamid

Nicholas Lewis

Introduction

Human minds are malleable, and prone to adopting various ‘radical’ ideologies. Some are harmful, others are harmless or even positive. Radicalism isn’t inherently negative, but it certainly can be inflammatory depending on the ideology which drives it and its proximity to violence.

In the deradicalization field, phrases such as ‘radical forgiveness’ and ‘radical compassion’ are used to express the counter-intuitive necessity of dropping prior assumptions about what it takes to effectively deal with severe mental health issues, extreme social polarization or isolation and various forms of hateful or controversial speech. There are numerous examples of the catharsis of facing a perceived enemy, from a child’s murderer in prison to a KKK member, in order to better understand them and emotionally process trauma.

This reality is on display across global social media platforms where the fabric of society appears to be fraying rapidly. Ironically, there seems to be a common belief across the political divide that Big Tech is mishandling their tremendous responsibility to maintain healthy dialogue. Some critics call for more censorship as an antidote for preventing the spread of harmful content. Others say free speech is a central requirement of civil discourse.

This report introduces data on the effects of deplatforming, which can enable both sides to form opinions based on long-term empirical research as opposed to short-term emotion.

The research found significant evidence that censorship and deplatforming can promote and amplify, rather than suppress, cognitive radicalization and even violent extremism. Shutting down accounts accused of violating hate-speech policies and misinformation often shifts those banned individuals to alternative platforms where their narrative of long-suffering victimhood is further refined.

The result is the creation of a framework for a new moderation system designed to elevate global discourse through Internet freedom. This was developed by experts in deradicalization, counter-extremism, communication, statistical analysis and social networking technologies.

This framework allows everyone to exchange and debate ideas civilly, and encourage other social media platforms to adopt this framework and join in our mission to reduce polarization, increase access to information and build a more healthy society.

One of the primary authors of this paper, Jesse Morton, tragically passed away early in 2022 before the release of this, his final work. He was the founder of Parallel Networks and research coordinator for the Institute for Strategic Dialogue's Against Violent Extremism Network in North America. Formerly a prominent radicalizer in the West, Morton co-founded and was chief propagandist of Revolution Muslim. His life is undeniable proof of the possibility of changing minds.

We are also led by Daryl Davis, a musician who, through communication and friendship, has helped more than 200 members of the Ku Klux Klan remove their hoods and permanently leave the group.

Our advisors include academics and technologists from around the world. Our goal is to use their expertise, experience and methods to better promote an open dialogue among world authorities about social media networks and conversations. We lay out our plan to accomplish these objectives, which are built around key goals:

- Take a critical look at current moderation methods, policies, and tools used by the large social media platforms
- Understand how radicalized individuals are spurred to violent and extremist actions
- Develop a more effective content moderation model and technology on Minds which preserves digital rights and empowers community governance
- Increase rate of deradicalization and depolarization in comparison to big tech platforms
- Provide mental health resources and communication outlets for anyone in need

There is an undeniable power in social media but it is also filled with shortcomings. Is it possible to find common ground through civil discourse and break out of toxic online echo chambers? We believe so and facilitate engagement, free expression, communication, and active input from parties who promote non-violence, critical thinking, creativity, and empathetic engagement. We call this the Change Minds Initiative.

Table of Contents

Foreword: Honest Dialogue Beats Silencing Opponents

By Daryl Davis

Humans are a remarkably varied lot. We have skin colors of many hues and support diverse political parties. We worship a wide range of deities and have different sexual preferences.

But there is something we share: everyone wants to be heard and understood. We want others to listen to our ideas. This desire has long driven human nature, and cultures are shaped by the debate of ideas in the public marketplace.

In a bygone era, we voiced our opinions through letters to the editor. While some detractors wrote angry missives, by the time it was mailed, received, and printed, that anger had typically subsided. A nuclear war of words was hardly imminent. These days, however, it seems like everyone on social media carries the nuclear codes, particularly influencers, and their finger is on the launch button. With a few keyboard strokes, we are on the path to either divisiveness or unification.

In response, we increasingly see the authors of controversial views suspended from, kicked off, or put in a platform's "jail" for expressing opinions. Exiled dissidents may seek alternate and sometimes nefarious platforms. Thus is born a breeding ground for conspiracy theories and plots. It's an ugly spiraling mess.

That's where the Change Minds Initiative comes in. We propose a methodology that will provide a space for all opinions hosted on Minds, a social media platform and humanitarian advocate.

The Change Minds Initiative is culled from multidisciplinary researchers and those with 'boots on the ground' experience, including former extremists, trolls, peacebuilders, activists, artists, and other professionals. All have been successful in deradicalization and the development of

communication skills that allow adversaries to exchange ideas in a civil fashion while leaving everyone's dignity and respect intact.

Change Minds is a team of diverse thinkers from varied backgrounds and perspectives. They have a simple but important goal: create a methodology that allows for inclusive dialogue where everyone can express their beliefs through thought-provoking conversation and debate. Change Minds is built on the belief that honest dialogue, not silencing opponents, remains key to conflict resolution. I am happy to join the team and participate in their Change Minds Initiative because I believe in inclusive dialogue and the promulgation of free speech, especially for those with whom you disagree.

We all have a moral imperative to combat darkness with light, to battle hate with love, and to counter extremism with compassion. Only then can we coexist and progress.

1.1 Content Moderation, Radicalization, and Social Media: Whack-a-Mole and Artificial Intelligence

"The phenomenon often referred to as 'incitement to radicalization towards violent extremism' has grown in recent years. This is mainly in relation to the Internet in general and social media in particular. This is despite it being immediately evident that other offline factors, including face-to-face communications, peer pressure and false information constitute more powerful forces, and are ignored at the peril of limiting our rights to freedom of expression if we focus only on the Internet."

- *Report from UNESCO, 2017*

Have you ever played the popular but twisted arcade game "Whack-a-Mole"? In it, several very chipper moles randomly pop up from holes and quickly disappear. You must bonk them with a wooden hammer before they zip away. If you've played the game, you might understand the frustration that many social media platform moderators, and policymakers, face in moderating unsettling content. They, too, deal with the pressure to quickly knock out something abusive that seems to come out of nowhere — then quickly disappears.

Historically, the moderation of online content is largely left to administrators and moderators of public forums, chatrooms, and other communication apps. Enforcement follows a perceived violation of a platform's community standards or Terms of Service. This purely human-based approach is inefficient, particularly on popular social media platforms, where massive amounts of content are produced — Facebook, Twitter, Reddit and other major social networks using closed-source software, have billions of active users daily. It's impossible for a single team to monitor it all. For years content moderation relied on underpaid humans who often labored under traumatizing (Newton, 2019) and harmful (Roberts, 2018) conditions for wages as low as $6/day (Vengattil & Dave, 2019).

Now, large-scale platforms increasingly rely on machine learning and artificial intelligence (AI) systems (Facebook, 2021b) as the first line of defense, often taking actions without contextual checks from humans, causing vast amounts of collateral damage in the form of mistakenly banned accounts. They perpetually crawl through the oceans of content in order to detect media that violates the platform's community standards, a number of which are outlined in the appendix of this paper.

There are several problems with this AI-reliant approach:

- It is difficult to develop nuanced systems that remove only content that violates the platform's rules while not over-censoring other material.
- AI is ultimately charged with upholding the platform's community standards. This can result in wrongly banned users who are unable to connect with a human moderator to appeal their case.
- The community has no role or voice in the process.

Typically, AI systems frame violations as classification problems: they classify content as either in violation of a policy (or not) and remove the content (or don't). Potential red flags often come via image recognition software designed to detect nudity — Facebook says it removes 99.2% of nudity through algorithmic classification (Taylor, 2020) — or in the form of terrorist content, where Facebook claims that its AI identifies 99.5% of terrorist content (Facebook, 2020). Said The New York Times: "Facebook's AI found 99.5% of terrorist content on the site, leading to the removal of roughly 1.9 million pieces of content in the first quarter" (Frenkel, 2018). The BBC added: "(Facebook) said its tools spotted 99.5% of propaganda posted in support of Islamic State, Al-Qaeda and other affiliated groups" (Lee, 2018). The impressive statistic says nothing about the volume of terrorist content on the platform, only that the AI found and flagged 99.5%.of terrorism content "before users reported it" (Magid, 2018). Such misrepresentations further the public's belief in the ability of AI to censor problematic content.

Current social media content moderation policies combat extremism by banning those who break their policies for posting hate speech and misinformation or violating their particular definition of

them. Yet, “hate speech” is often used as an umbrella term that depends on the mood, political persuasion, or opinion of human content moderators (York & McSherry, 2019). “Hate” is part emotion, part chronic negative disposition of a person or group (McCauley, 2020), and “hate speech” is an abstraction of that emotion. Defining what constitutes “hate” or “hate speech” can be subjective and elusive (Fino, 2020), and attempting to ban “extreme” speech will almost always serve as a path into censorship. It’s a slippery slope on which a small first step leads to a tumbling chain of events that culminate in counterproductive effects and unintended consequences (Jozwiak, 2018).

For example, more than 20 years after the horrific attacks of September 11, 2001 and the beginning of the War on Terror, legislative actions that curtailed civil liberties, chilled dissent, and failed to effectively address the root causes of terrorism. There are at least four times as many Sunni Islamic militants today as there were on 9/11 (Jones et al., 2018), suggesting that our response to terrorism is not only been counterproductive but caused terrorism to grow. Calls to wage a domestic War on Terror (Shuster, 2021) in the aftermath of the January 6, 2021 storming of the U.S. Capitol follow a similar trajectory. Powerful social media companies tighten their content moderation policies to remove hate speech, while public experts and authorities presume a direct documented link between speech and violent extremism. Such claims are dubious, and that is the central question we explore in this paper: Is the link between social media usage, hate speech, and violent extremist action supported by sound scholarly evidence? Do current methods of content moderation lead to further toxic partisan polarization, radicalization and dehumanization (Coleman, 2021)? An August 2021 Pew Research Center poll found that 59% of U.S. adults say those tech companies should restrict false information, even if it means losing some freedom to access and publish content. These numbers are divided along political lines, with 76% of Democrats and 37% of Republicans holding such opinions (Mitchell & Walker, 2021).

Gallup polls conducted after the January 6, 2021 riots in Washington, D.C. showed that 51% of liberals hold negative views of Big Tech platforms for what they say is a failure to take strong action to curtail hate speech and disinformation. Conservatives accuse social media companies of censoring right-leaning content (Shepardson, 2019). Those feelings are growing. Over 18 months, the percentage of Republicans who had positive views of Big Tech firms fell from 43% to 2%, while support from independents in the Gallup survey fell from 43% to 33%. (Brennan, 2021). Popular conservative politicians such as former president Donald Trump and Senator Ted Cruz

(R-TX) as well as right-wing commentators such as Fox News personality Tucker Carlson helped further spread the notion that Big Tech censors conservative views (Rampton & Shepardson, 2019). Minds CEO Bill Ottman appeared on both National Public Radioand “Tucker Carlson Tonight” , saying that Big Tech censors both the left and the right.

This confluence of factors creates a hyperpolarized social media landscape, with the increased potential for mass future migration to alternative platforms (Wong & Morse, 2021), many promoting less restrictive content moderation policies.

Minds systematically studies the most efficient, ethical, and innovative methods for combating hate and violent extremism on social media platforms while avoiding harms that can come from over censoring which can stunt psychological growth, increase the sense of entitlement, decrease control of emotions and outbursts, enhance aggression, and delay the development of coping skills. The objective is to create a content moderation model that preserves free expression, promotes engagement, and utilizes evidence-based interventions to promote the disengagement, deradicalization and re-humanization of at-risk individuals and collectives.

This approach will be in line with the current understanding of online influence into radical extremist networks. For example, a Facebook internal report from 2016 documented that 64% of people who joined an extremist group on the platform did so because the company’s algorithm recommended it (Bauerlein & Jeffery, 2021). According to The Wall Street Journal, a 2018 presentation at Facebook included a slide that read: “Our algorithms exploit the human brain’s attraction to divisiveness.” It also included a dire warning. “If left unchecked,” it warned, Facebook would feed users “more and more divisive content in an effort to gain user attention & increase time on the platform.”

After adjusting the algorithm, Facebook shelved the alterations. People spent less time on the platform and Facebook chose to increase, rather than decrease, the polarizing effects because such addictiveness is profitable (Horwitz & Seetharaman, 2020). Facebook is not alone. Social media networks, including YouTube and Twitter, are criticized because of their algorithms for promoting content on user feeds to increase the time spent on the platform. It’s reasonable for a company to want customers to use their product, but to achieve this by manipulating the user

without any transparency is problematic. Because of this, many argue that AI promotes extremism, a view expressed in the 2021 documentary "The Social Dilemma," for example.[1]

The words extremist and radical are grossly unequipped to properly communicate the vast spectrum of interpretations when we hear them. Much of their use is subjective and becomes diluted when used improperly or as part of *ad hominem* arguments without sufficient evidence. As Supreme Court Associate Justice John Marshall Harlan stated, "One man's vulgarity is another man's lyric". This came in his majority opinion in the Cohen vs. California case which protected a Vietnam War protestor wearing a jacket bearing the words "Fuck the Draft". But it can just as easily be applied to social media.

[1] Available at https://www.netflix.com/title/81254224.

1.2 Free Speech vs. Hate Speech

“You can’t change someone’s mind if you don’t give them a platform to speak.”

Adopted in 1791, the First Amendment of the Bill of Rights in the U.S. Constitution expresses the fundamental principle that the government “shall make no law... abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.” So-called “extremist speech” is generally protected under the First Amendment as interpreted by the Supreme Court unless those words can be defined as a “true threat,” which would entail a statement meant to frighten or intimidate one or more specified persons into believing that they will be seriously harmed by the speaker or by someone acting at the speaker’s behest (O’Neill, 2009).

Today, the individual right of free speech and expression is enshrined in the United Nations Declaration of Human Rights, and most constitutions in the world with some limitations — European countries banned “extremist” speech, such as Holocaust denial. Supreme Court Justice Oliver Wendell Holmes, Jr. also famously wrote, “The most stringent protection of free speech would not protect a man falsely shouting fire in a theater and causing a panic..... “

Shutting down dialogue and engagement with those espousing intolerant perspectives can increase the rate of radicalization. Consequently, such speech is driven further underground, which has created a martyr complex amongst some adherents to such perspectives (Nicas, 2018; Paul, 2019).

This is an important moment to reassert the right to free speech. With the power of social media in our everyday lives, it is essential to recognize Big Tech’s power to effectively determine and regulate what is considered hate speech. That’s an awesome responsibility. Today, private companies may be more influential than democratic governments in determining the boundaries of free expression. In Packingham v. North Carolina (2016), for example, Supreme Court Justice Anthony Kennedy noted: “While in the past there may have been difficulty in identifying the most

important places for the exchange of views today the answer is clear. It is cyberspace — the 'vast democratic forums of the Internet' in general, and social media in particular."

The content moderation policies of major social media platforms must be grounded in empirical evidence that are shown to be the most effective approach at minimizing violent extremism. They should be based on an objective, consistent and evidence-based approach. The status quo across Big Tech prohibits activity proven to counter violent extremism. You can't change someone's mind if you don't give them a platform to speak it. A focus on countering internal biases (e.g., moderator opinions), transparency, and targeting methods of influence (e.g., trust-building, friendship) that may lead those who are amenable to alternatives to shun networks with less access to diverse opinions.

Corporate definitions of what constitutes hate speech, however, are more problematic. First, there is little legal guidance on the matter. There is no legal definition of hate speech under U.S. law — as opposed to what constitutes a hate crime — just as there is no legal definition for evil ideas, rudeness, unpatriotic speech, or any other kind of speech that people might condemn (Ward, 1997).

Social networks define the boundaries for acceptable content and what they judge to be hate speech (see Appendix: Definition of "Hate Speech" on Large-Scale Social Media Platforms). But these guidelines are ambiguous and have faced scrutiny (York & McSherry, 2019) for a lack of transparency (Singh, 2019) and inconsistency in the application of moderation policies (Langvardt, 2018; Tworek, 2021). That's a huge problem. What deserves to be censored isn't always clear, to both AI systems and human moderators.

This is further complicated because hate speech in one cultural context may not be considered offensive in another (example: the use of the N-Word) There may also be generational divides; what is considered to be offensive now may not always have been. With such a mixed and nuanced bag of definitions and guidelines, it's no wonder that human moderators or algorithms struggle, often revealing more about the person or platform defining the concept than the actual effect of the content.

Large-scale social media platforms say they censor content that promotes violence on groups with protected characteristics such as race, religion, sexual orientation, and gender. But there are serious reasons to be wary of powerful tech companies setting the rules for what is permissible to say in a democratic society, especially in the U.S. where such power can supersede the Constitution's First Amendment. As Peggy Hicks, Director of Thematic Engagement at the United Nations Human Rights Office explained in July 2021: "We have the same rights online as offline. But look at the online landscape, and you see a digital world that is unwelcoming and frequently unsafe for people trying to exercise their rights. You also see a host of government and company responses that risk making the situation worse…" Supreme Court Justice Clarence Thomas has openly advocated for networks of a certain size to be deemed Common Carriers, akin to phone companies which are not able to discriminate against customers.

Facebook, in particular, is heavily criticized for carrying out the censorship objectives of governments such as Russia, India, Turkey, South Korea, Brazil, Vietnam and Thailand, taking down posts and banning pro-democracy and human rights users at governments' request in exchange for maintaining access to those countries' markets (Clarke & Swindells, 2021). YouTube's algorithms removed content that can be central to the prosecution of war crimes and other international infractions in a manner that makes the material inaccessible to investigators and researchers holding perpetrators to account (Wille, 2020). Twitter deleted user accounts that highlighted the eviction of Palestinians from East Jerusalem. Said Jillian York, director for international freedom of expression at the Electronic Frontier Foundation: "Moderation is on the rise, and it's really a blunt object...the companies don't pay enough attention to cultural contexts such as Palestine where there's basically less profit, so they put a lot more effort into making content moderation and automation effective in larger markets. (Gebeily, 2021)."

These large-scale platforms cultivate the image of a constantly improving, consistent, rule-based, big data-driven content moderation approach. But that conceals a range of inconsistent decisions and false positives. How can improvement be measured when the key performance indicators are diametrically opposed depending on whether censorship of lawful speech is supported or not? While Big Tech ostensibly frames content moderation as the language of human rights, their actions seem more driven by profit and business, the risk of governmental regulation, outside pressure from the public, media pressure, and the partisan academic and think tank community.

Microsoft, the owners of LinkedIn, have placated the Chinese Communist Party by censoring journalists critical of the government there, while that government used the platform for espionage and recruitment. Meanwhile, in the run-up to recent parliamentary elections in Russia, Google and Apple complied with Russian government demands to block access to content relating to Alexia Navalny, the pro-democracy, human rights activist poisoned by Russian agents in August 2020 and now in a Russian prison. Both platforms were used by Navalny's allies to coordinate voting (Troianovski & Nechepurenko, 2021).

Perhaps the most egregious offender is Facebook. In September 2021, the Wall Street Journal released a series of articles called the Facebook Files which were based on a review of internal Facebook documents, including research reports, online employee discussions and drafts of presentations to senior management. Time and again, the documents show, Facebook's researchers have identified the platform's ill effects.

The Facebook Files were discussed in February 2022 directly with Mark Zuckerberg on episode 267 of "Lex Fridman Podcast". Zuckerberg claimed that Facebook spends more money than most social media companies on research, and cares about issues pertaining to mental health and free expression deeply. If this is the case, then it is bewildering why so much of the research and source code is kept secret which alleges to be addressing the issue. It is equally a quandary why none of the research that appears in this paper and elsewhere, about the blowback and inefficacy of censorship, has ever been acknowledged openly. If a full-scope academic analysis was taking place, where is the proof that all relevant data has even been considered in order to form a highly informed opinion?

Facebook's misguided approach is evident in how the platform addresses other negative activities, including a so-called "White List" that exempts certain high-profile businesses and celebrities from some or all of its rules. An internal document reads, "For a select few members of our community, we are not enforcing our policies and standards. Unlike the rest of our community, these people can violate our standards without any consequences (Horwitz, 2021)."

For years, researchers on Instagram, which is owned by Facebook (now Meta Platforms), informed higher-ups that the photo-sharing app on the platform is toxic for teen girls. Facebook did little to address these concerns despite contrary statements to legislators and an

unwillingness to release internal data to external researchers (Wells, Horwitz & Seetharaman, 2021). Employees raised concerns about how Facebook is used nefariously in developing countries —for example, human traffickers in the Middle East, and repressive government action against political dissent. But Facebook's internal documents show that in many instances the company did nothing to quell these abuses. (Scheck, Purnell & Horwitz, 2021).

It is unreasonable to believe that it is a social media app's primary responsibility to maintain the mental health of all users or that mental illness can be eradicated. However, the complete lack of transparency from Facebook and others. about the logic of their recommendations algorithms is grounds for an inquiry around malicious intent, which can't be disproven without disclosing the source code.

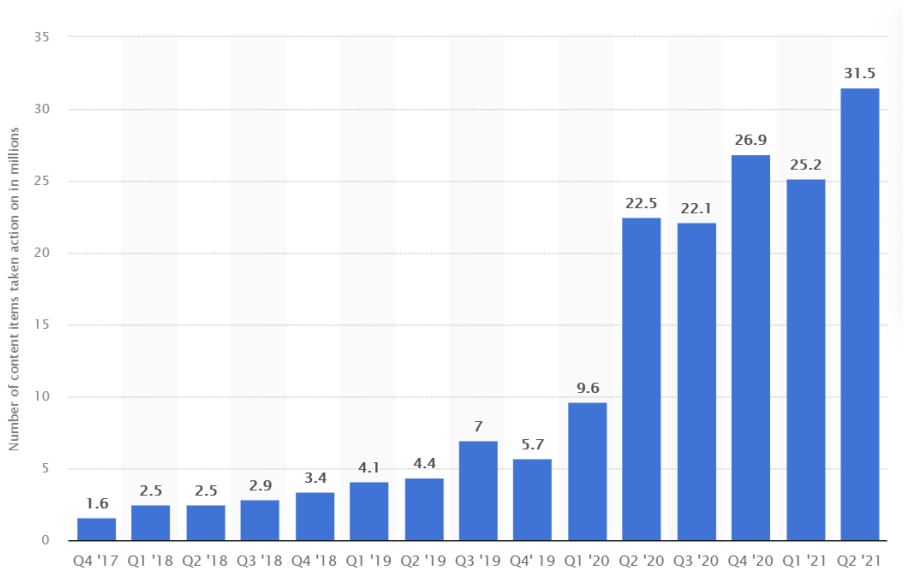
These are not the only offenders. Other large-scale platforms are driven by similar motives, so it is unlikely that their profit-over-people business calculus differs behind closed doors. Yet, within larger markets, the risk of a more hostile and costly regulatory framework and enhanced public pressure have induced a lot more effort to make content moderation a priority. Developments in AI detection and exaggerated claims about AI's capability to moderate initiated another slippery slope of censorship. What began as earnest efforts to use AI to identify and remove violent extremist propaganda from organizations like ISIS and Al-Qaeda has broadened to remove content that may be radical or even hateful but certainly refrains from calling to violence. There is no distinction between these in Big Tech policies, which lumps together dark comedy and violent extremism.

During the campaign preceding the U.S. 2020 presidential election, Facebook's AI disabled tens of thousands of accounts related to conservative organizations and perspectives, flagging them as spam or hate speech. Many of these were falsely flagged (Volz, 2020). After the illegal actions of protestors at the US Capitol and under the belief that it was a pre-planned insurrection, it was reported that Facebook videos even featuring the voice of former President Donald Trump were removed for violations of community standards (BBC, 2021). Former President Trump remains restricted from Facebook despite their oversight board (Culliford, 2019) saying, "the company was wrong to impose an indefinite ban" (Bond, 2021). Trump's ban is active while accounts from world leaders like Iran's supreme leader Ayatollah Ali Khamenei, with frequent calls for 'death to America' (Morrison, 2021), remained on the platform.

Twitter is similar. Despite Twitter's ex-CEO Jack Dorsey tweeting that, "over the long term [Trump's ban] will be destructive to the noble purpose and ideals of the open internet. A company making a business decision to moderate itself is different from a government removing access, yet can feel much the same," Donald Trump's account is still "permanently suspended (...) due to the risk of further incitement of violence" (Twitter Inc., 2021). Yet, the Taliban currently operate with impunity on the platform and has more than 800,000 followers (Chibelusi, 2021).

Today, the content moderation policies of large-scale social media companies blur any clear distinction between hate speech and content that promotes violent extremism. Speech such as "shitposting" (Haque, 2019; McEwan, 2017), the activity of posting deliberately provocative or off-topic comments on social media, or regular communications about incidents of violence, can easily trigger censors of hate speech for removal and banning (Dodgson, 2019). While censorship practices negatively affected all kinds of people, they particularly impacted marginalized communities - not only working-class conservatives and conspiracy theorists (Abril 2021), but also communities of color, women, LGBTQ+ communities, and religious minorities (Díaz & Hecht-Felella, 2021).

The slippery slope of censorship keeps getting tricker to negotiate while information on social networks is less reliable. As the chart below shows, in the fourth quarter of 2017 Facebook removed 1.6 million pieces of hate-speech content. This was as the platform faced a firestorm of criticism for its role in the 2016 election, and in the immediate aftermath of CEO Mark Zuckerberg's admitting that Facebook was not prepared for foreign actors (Benkler, Faris, & Roberts, 2018). Less than four and a half years later, however, the number of removed pieces of content has risen to 31.5 million per year (see chart below).



If censorship was working then these trends should be holding steady or decreasing, as their algorithms flag more content. But it appears that hate speech and misinformation are at an all-time high, suggesting that their own approach to censorship doesn't work. This indicates a troubling new reality of over-regulating content and jeopardizing the universal right of free expression. Given the large number of false positives regarding hate speech, and confusion in the automated and manual moderation of content, important questions arise: Have the policies of social media companies regarding censorship prevented violence? Does silencing problematic voices decrease violence on these platforms and in real life?

1.3 A Critical Review of the Evidence

"The exposure-causing-radicalization hypothesis fails to explain the most likely outcome of exposure to radical material: rejection."

Calls for censoring hate speech and violent extremist content on social media platforms are common (Kimball, 2019), but there's little evidence that links exposure to extremist material online and "incitement to radicalization towards violent extremism" (UNESCO, 2017). There is a presumed monocausal direct link between 'hate-speech' online and the commission of violent extremism — a sentiment that empowers those calling for more aggressive social media censorship to prevent or counter violent extremism.[2] It also galvanizes extremists, who interpret overly simplistic media, commentary, and censorship to confirm their narrative of victimhood (Kaufman, 2020) or conspiratorial thinking (Buckley, 2015).

A systematic review of the academic literature focused on the link between extremist online content and violent radicalization, however, finds little evidence of a quantifiable connection. The study "Exposure to Extremist Online Content Could Lead to Violent Radicalization: A Systematic Review of Empirical Evidence", which aggregates more than 5,000 other studies on the topic, found that only 11 included "tentative evidence that exposure to radical violent online material is associated with extremist online and offline attitudes, as well as the risk of committing political violence among white supremacist, neo-Nazi, and radical Islamist groups" (Hassan et al., 2018).

But the authors said none of the 11 studies conducted a meta-analysis "due to the heterogeneous and at times incomparable nature of the data." But one of the takeaways was that active seekers, as compared to passive receivers, of violent radical material seemed to be at higher risk of engaging in political violence. If that is the case, then restricting extremist content from

[2] The term *countering violent extremism* (CVE) is defined by the U.S. Department of Homeland Security as all proactive actions taken to "*counter efforts by extremists to recruit, radicalize, and mobilize followers to violence*." All in all, it encompasses all activities undertaken to address the underlying conditions and root factors that can contribute to recruitment and radicalization by violent extremists. For more information, see: https://www.dhs.gov/cve/what-is-cve

large-scale social media platforms is unlikely to prevent those actively pursuing it and more likely to radicalize them toward violence.

In 2013, RAND Corporation released a study that explored how the internet was used by 15 convicted violent extremists and terrorists during their process of radicalization.

The report utilized evidence presented at trial, computer registries of convicted terrorists and interviews with convicted terrorists and senior investigative officers responsible for terrorist investigations to test five hypotheses generated by a review of the existing literature. These were: 1. the internet creates more opportunities to become radicalized; 2. the internet acts as an 'echo chamber,' a place where individuals find their ideas supported and echoed by other like-minded individuals; 3. the internet accelerates the process of radicalization; 4. the internet allows radicalization to occur without physical contact; and, 5. the internet increases opportunities for self-radicalization. The findings confirmed that the internet played a small role in the radicalization process of the 15 violent extremists and terrorists studied and supported the suggestion that the internet may act as an echo chamber and enhance opportunities to become radicalized. Yet the evidence did "not necessarily support the suggestion that the internet accelerates radicalization, nor that the internet allows radicalization to occur without physical contact, nor that the internet increases opportunities for self-radicalization as in all the cases reviewed during, the subjects had contact with other individuals, whether virtually or physically" (Von Behr, 2013).

The limited empirical evidence that exists on the role that online speech plays in the radicalization-to-violence journey suggests that people are primarily radicalized through experienced disaffection (Hafez & Mullins, 2015), face-to-face encounters and offline relationships (Dreyfuss, 2017). Extremist propaganda alone does not turn individuals to violence (Gilsinan, 2015). Other variables are at play.

Another analysis compiled every known jihadist attack (successful & thwarted) from 2014-2021 in 8 western countries with 439 cases across 245 attacks/plots. "Our findings show that the primary threat still comes from those who have been radicalised offline. [They] are greater in number, better at evading detection by security officials, more likely to complete a terrorist attack successfully and more deadly when they do so" (Ariza & Hamid, 2022).

Research with convicted former terrorists shows that media coverage of radicalized groups inspires them to find alternative narratives online. Consequential exposure to propaganda that blamed the media for the societal rejection of their ingroup cemented that interest (Baugut & Neumann, 2020). More recent research into the social media usage of convicted terrorists in the United Kingdom found that those who were primarily radicalized online were less socially connected and less identified with an extremist group or cause. They were thought least willing and able to perpetrate violent extremist acts (Kenyon, Binder & Baker-Beall, 2021).

In other words, the focus on social media usage as a radicalization vector may be overstated. More attention on the role mainstream news coverage through TV and print media may be required. Instead, the media frequently covers social media usage as a gateway to extremist brainwashing. This is despite the fact that most people shown extreme content will reject it. Very few will act upon it in a violent manner. As such, there must be more focus on the process of radicalization; otherwise, the exposure-causes-radicalization hypothesis fails to explain the most likely outcome of exposure to radical material: rejection.

Even less scientific rigor has been applied to researching the ultimate consequence of removing hate speech and misinformation from social media platforms. We will not see the true impact of January 2021's social media banning and censorship spree on platforms such as Facebook, Google, and Twitter for a while. These efforts may even make matters worse. A study from George Washington University in 2019 found that: "[T]he key to understanding the resilience of online hate lies in its global network-of-network dynamics. Interconnected hate clusters form global 'hate highways' that—assisted by collective online adaptations—cross social media platforms, sometimes using 'back doors' even after being banned, as well as jumping between countries, continents and languages" (Johnson et al., 2019).

Online, extremists are adept at migrating to obscure and difficult-to-monitor social media platforms. As they do they adjust the message and mechanisms of propaganda dissemination, distribution and recruitment. In other words, banning words and memes causes the memes to mutate. The activity of mutating memes and symbols is popular, and even welcomed, among online trolls.

A more recent analysis by the same team at George Washington University analyzed hate speech and COVID-19 misinformation for the white supremacist movement and “medical disinformants” and mapped these movements at an ecosystemic level. They reviewed six platforms that included both mainstream and alternative social media networks: Facebook, Instagram, Gab, 4chan, Telegram, and VKontakte. Researchers looked at the hyperlinks between clusters and found that harmful content, including hateful posts and COVID-19 misinformation narratives, spreads quickly between platforms.

Hyperlinks facilitate this by acting like “wormholes” that transport users between platforms in a click that crosses space, time, and moderation regimes, suggesting that users move between moderated and unmoderated platforms. As the authors note: “An extremist group has incentives to maintain a presence on a mainstream platform (e.g., Facebook Page) where it shares incendiary news stories and provocative memes to draw in new followers. Then once they have built interest and gained the trust of those new followers, the most active members and page administrators direct the vetted potential recruits towards their accounts in less-moderated platforms such as Telegram and Gab, where they can connect among themselves and more openly discuss hateful and extreme ideologies” (Velásquez et al., 2021).

This funneling behavior is an expected result of network migration techniques adapting to censorship policies. But how many violent extremists or borderline violent extremist individuals have Big Tech companies helped? How many have they engaged or aided in deradicalization?

On a few occasions, Minds has been accused of propelling extremist content alongside other companies in the emerging industry of alternative and decentralized social networks. This criticism comes despite an evidence-based approach to these problems. The inconvenient truth of online hate and misinformation is that solving, not hiding, requires long-term communication with these individuals.

There is not a major social network or platform that is effectively solving this problem, so why not A/B test a different long-term approach? As computer scientists, should this not be ground zero for determining policy in the first place?

This migration of information from large-scale platforms (where feedback from a general population may counter and moderate extreme views) to smaller platforms (where extreme beliefs can be reinforced by others who have also been banned by larger platforms) creates powerful echo chambers where controversial ideas are reinforced and their wording carefully rephrased before they are reintroduced to larger social media platforms to avoid content violations. Additionally, propagandists post new inter-platform links to the smaller scale platforms. In the long run, the number of those in the extremist echo chambers on smaller scale platforms grows. Enhanced efforts to police large-scale platforms promotes further narratives of censorship and victimhood, increases commitment and, thereby, the likelihood of extremist support, engagement and radicalization to violence. As Erin Saltman, formerly Facebook's head of counterterrorism and dangerous organizations policy for Europe, the Middle East and Africa, expressed: "Online terrorism and violent extremism are cross-platform and transnational by nature. Nobody has just one app on their phone or their laptop, and bad actors are no different, [thus] any efforts trying to effectively counter-terrorism and violent extremism need to similarly go beyond one-country, one-platform frameworks." Saltman (2021) calls this the "next big challenge" for governments and nongovernmental organizations (NGOs) working with tech companies.

The shift in the mainstream social media companies' concentration on content removal can be traced back to the rise of ISIS and its caliphate in June 2014. A seminal Brookings study, "The ISIS Twitter Census" (Berger & Morgan, 2015), analyzed the online presence of ISIS and supporters at a time when their savvy use of social media, most notoriously on Twitter, shocked the world. The study found between 46,000-70,000 active pro-ISIS Twitter accounts from September through December 2014. Soon thereafter, Twitter suspended a large number of ISIS-supporting accounts.

This blunt approach could produce ominous ramifications. The report stressed that "further study is required to evaluate the unintended consequences of suspension campaigns and their attendant trade-offs. Fundamentally, tampering with social networks is a form of social engineering," and that "the process of suspension does create certain new risks. Most importantly, while suspensions appear to have created obstacles to supporters joining ISIS's social network, they also isolate ISIS supporters online. This could increase the speed and intensity of radicalization for those who manage to enter the network, and hinder organic social pressures that could lead to deradicalization."

The authors concluded: "Social media platforms should consider whether they want to continue with some variation of their current approach, which tends to stomp out fires as they erupt, or whether they want to dismantle or degrade the social networks responsible for setting the fires." They also addressed the legislative elephant in the room: "It is unwise for social media companies to presume they will remain immune to regulation. Companies should get out ahead of the curve by crafting policies and publicly articulating their priorities. If they do not bring their vision to the government, the government is likely to bring a much more restrictive vision to them."

The paper, released in March 2015, garnered significant media coverage (Gladstone & Goel, 2015) and Twitter responded without addressing the warnings of unintended consequences and the need for further study.[3] Instead, Twitter accelerated its efforts. "As the nature of the terrorist threat has changed, so has our ongoing work in this area," the company said in February 2016. "Since the middle of 2015 alone, we've suspended over 125,000 accounts for threatening or promoting terrorist acts, primarily related to ISIS" (Twitter, 2016).

The same authors of the Brookings study released a paper in February 2016 about the potentially negative impact of deplatforming. Titled "The Islamic State's Diminishing Returns on Twitter: How Suspensions are Limiting the Social Networks of English-speaking ISIS supporters," the report analyzed the metrics of a network of English-language ISIS supporters active on Twitter from June to October 2015 and proclaimed that "suspensions held the size and reach of the overall network flat, while devastating the reach of specific users who have been repeatedly targeted" (Berger & Perez, 2016).

The report also described the effects of extremist adaptation: "In recent months, Telegram Messenger has emerged as a favored alternative to Twitter for the initial publication and dissemination of official Islamic State propaganda." Telegram, a private messaging platform, is almost impossible for law enforcement agents to monitor. In November 2015, as Twitter increased its account suspension efforts, directly causing a cell of ISIS affiliates to use Telegram to communicate as they killed 130 people in a series of coordinated attacks across Paris in the deadliest jihadist attack in French history (Alexander & Braniff, 2018). ISIS inspired or directed

[3] As the paper stressed, "It is unwise for social media companies to presume they will remain immune to regulation. Companies should get out ahead of the curve by crafting policies and publicly articulating their priorities. If they do not bring their vision to the government, the government is likely to bring a much more restrictive vision to them."

additional attacks in Europe using the application, while supporters commented prolifically on Twitter's censorship (CBS News, 2019). The attacks included a December 2016 truck-ramming attack on a Christmas market in Berlin, a mass shooting at the Reina nightclub in Istanbul (Shehabat, Mitew & Alzoubi, 2017), and a spate of vehicular attacks in Europe and the U.S. (Counter Extremism Project, 2020) carried out by disgruntled ISIS supporters encouraged and instructed on Telegram (Clifford, 2018).

Ominously, the report noted that similar content moderation recommendations may have a similar effect on far-right wing extremism. Large-scale platforms may be repeating similar mistakes they made with regard to ISIS byapplying them to domestic entities and individuals who currently do not call for violent extremism or terrorism. Millions of conservatives migrated to Telegram (Nicas, Isaac & Frenkel, 2021) after Twitter and Facebook shut down then-president Trump's accounts in a span of little more than 24 hours. This was followed by Amazon, Apple, and Google removing Parler, a social media site that grew by millions in the wake of the U.S. presidential election largely because it touted itself as a place to "speak freely and express yourself openly without fear of being deplatformed for your views" (Yurieff, Fung & O'Sullivan, 2021). Recent pressures to remove COVID-19 conspiracy theories, QAnon, the Proud Boys, the Boogaloo Boys, have forced those who support such ideologies (few of which call for violence) to adapt. After Facebook banned the word "boogaloo," for example, the group reorganized by adjusting its keywords and hashtags (Tech Transparency Project, 2020). QAnon supporters used memes and other visuals without captions, linking to smaller-scale platforms to attract a newer audience and maintain momentum. This while suggesting the ban was evidence that proved their conspiratorial perspective (Frenkel, 2020).

While Facebook, Google, and Twitter's AI now remove 72% of the hate speech that is illegal in the EU, according to the European Commission, only half of user-reported posts were removed (Carlson & Rouselle, 2020). In addition, during the COVID-19 pandemic, several experts, including Robert Malone, who was central to the discovery and development of mRNA vaccines, had videos removed from YouTube on the ground that they spread misinformation and conspiracy, despite his being pro-vaccine and involved in the ultimate development of the COVID-19 vaccines. Similar deplatforming was reported regarding the use of Ivermectin, a drug used as an antiparasitic agent (for which it received a Nobel Prize in 2015). Although its use has had mixed results in the

literature,[4] with some showing it to be ineffectual and others showing it to be effectual as a treatment for COVID-19 symptoms, social media sites have used censorship and deplatforming to stifle public debate. This has several chilling effects, including mistrust of treatments that have proven effective and the development of conspiracies as it is unclear why the debate is stifled during a public health crisis. These censorship events quickly become top trending global news as is expected from studies on the "Streisand Effect" which is described in more detail below. For instance, podcast interviews with Dr. Peter McCullough and Dr. Robert Malone on the "The Joe Rogan Experience" caused Spotify employees to revolt against the existence of the content on the service. These episodes became viral sensations discussing criticism of certain responses to the pandemic. Both Neil Young and Joni Mitchell had their music removed from the service as well. Ultimately the boycott failed to convince Spotify management but they did add more explicit COVID-19 guidelines alerts.

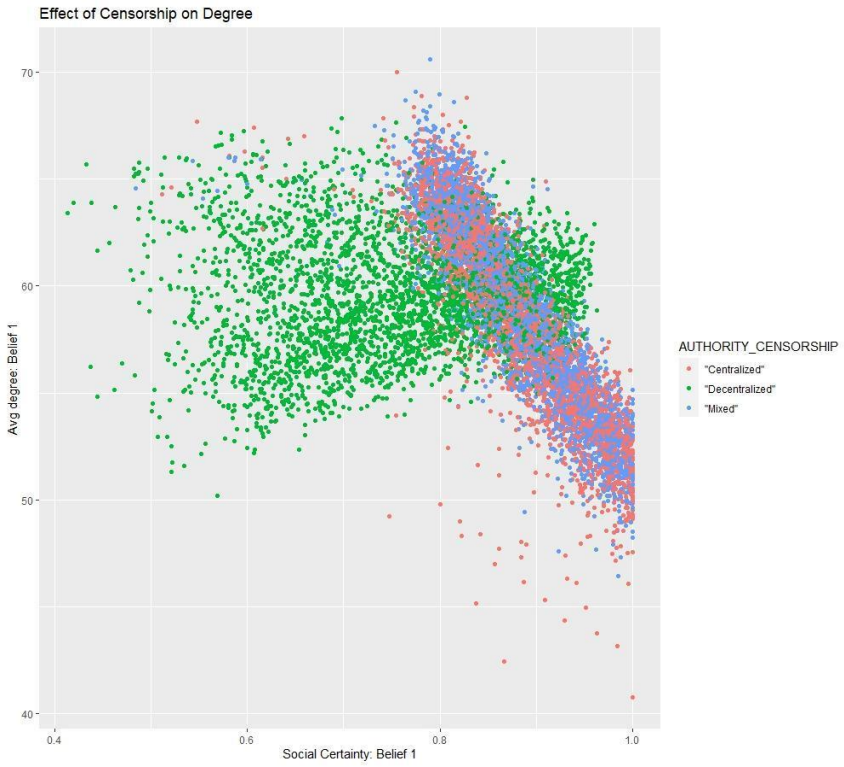
If the internet and social media are not the main means of radicalization — a complex confluence of other variables and face-to-face engagements serve as more predominant instigators (Dreyfuss, 2017) — then popular notions of online extremist recruitment and calls to censor "hateful" content must then also be questioned (McCauley & Moskalenko, 2010). According to the data, radicalized individuals are likely searching for online engagement that confirms their preexistent grievances. And today, if they seek it, they will certainly find it. The George Washington University study concluded much the same. Based on their observations and analysis, the authors developed a mathematical model that predicted, "policing within a single platform (such as Facebook) can make matters worse, and can generate global 'dark pools' in which online hate will flourish" (Johnson et al., 2019).

Censoring may also inflame those tip-toeing around radicalized passion. In a study completed by political scientists and computer simulation and AI experts (Lane, McCaffree, & Shults, 2021), the authors developed a computer model representing a social network. They then manipulated the

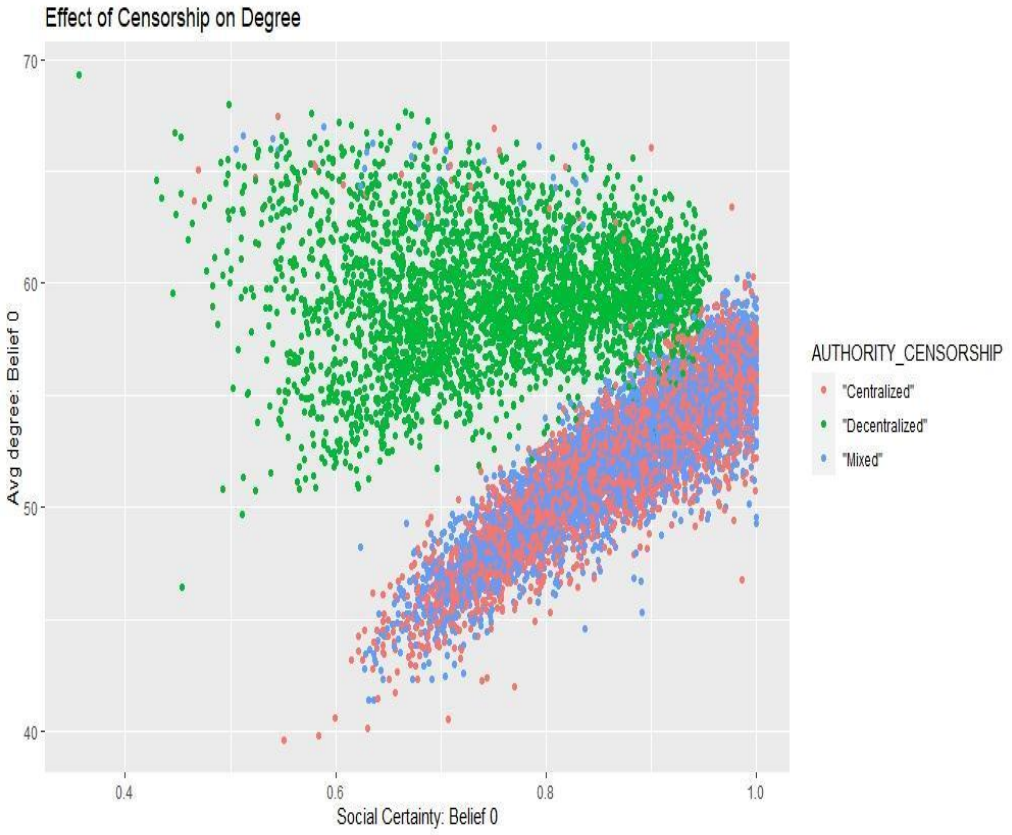
[4] For example, while papers published in some peer-reviewed journals report systematic reviews and meta-analyses of studies suggesting that overall its use lowers mortality risks (see Kow et al., 2021 and Padhy et al, 2020), other researchers have rightly entered the debate in critique suggesting that methods of underlying studies might be flawed and that more research is needed given the global health risks involved (e.g., Lawrence et al., 2021). While this is a logically sound critique, members of the general public are right to also question why the perceived skepticism is not met with equal fervor for more research on the topic, when such a great deal of energy, resources, and funding were poured into the development of the current vaccine options. From a public health communication standpoint, this is a valid concern that should have been more clearly addressed by medical professionals, or, at the very least, have been a debate that was allowed to take place in social media channels among members of the public.

mechanisms of content moderation so that, in some simulations, there was a globally defined censorship mechanism that could cause an agent to be banned and isolated from the network if they are found to hold a specific belief. In other simulations, there was a decentralization of content moderation mechanisms and all agents could block others so that no new information from blocked individuals was shown. These two methods were tested in conjunction, allowing for the exploration of mixed-method censorship mechanisms (like those used at Facebook) where the independent agents and a centralized authority could enforce bans. The difference here is that while a ban by a central authority affected the agent's ability to interact with anyone else on the network, an independent agent's ban still allows for interactions with others who haven't banned the agent with the belief in question.

These censorship protocols had wildly different effects. Allowing members to decentrally decide what they see dispersed the strength of belief in antisocial views and lowered the level of certainty in socially unacceptable beliefs. In centralized censorship models or in mixed models, there was effectively no difference in the certainty in socially unacceptable, or radical beliefs. But the simulation found that as agents become more isolated, on average, their certainty in their beliefs gets stronger (as shown in the graph below):



When looking at the effect of censorship for nonradical views, however, the simulations showed that when censorship is fully decentralized, agents' certainty of beliefs deemed socially unacceptable does not substantially change. For socially acceptable beliefs, on the other hand, individual agent certainty is reinforced socially, as evidenced by the positive relationship between certainty and degree centrality (number of links in the simulation). These effects are shown in the graph below. Thus, the modeling showed that decentralized censorship (that is: between individuals only) has little effect on the creation of radical individuals, but in fact appears to mitigate radicalization. The presence of any censorship mechanism, even mixed censorship mechanisms (such as those seen on Facebook and Twitter), encourages isolation in echo chambers. This results in greater certainty in socially unacceptable beliefs and radicalization. These findings are also backed up by other studies that did not use computational modeling (Ali et al., 2021)



In summary, there is not enough evidence to suggest that current censorship approaches, such as those used by the biggest social media companies, actually prevent negative real-world outcomes. On the contrary, expanding the scope of content moderation around hate speech might actually promote radicalization and violent extremism, particularly in the long run. Removing subjectively defined hate speech forces those who have been censored to revert to smaller platforms, where they become stuck in echo chambers in which their grievances can be galvanized, victimhood narratives affirmed, and violence encouraged as the only recourse.

1.4 A Case Study on the Fundamental Flaws in the Research Suggesting Deplatforming is Effective

Few of the studies discussed attempt to quantify the effects of censorship on user engagement, which measures the time a user spends on a page on your website. Instead, many studies in the area of countering violent extremism, hate speech and censorship focus on more subjective qualitative analyses — and these can be devalued by researcher bias. As an example, one widely cited paper of the effects of deplatforming on engagement was published in 2021 by the Berkman Klein Center for Internet and Society at Harvard University. It analyzed more than 11,000 YouTube channels in an attempt to measure the effects of censorship on YouTube channels such as Alex Jones (Kaiser, 2021), a controversial radio and online video pundit. Over multiple years, Jones became one of the most prominent global social media influencers, but in August 2018 YouTube and Facebook banned him for violations of community standards (Chappell & Tsioulcas, 2018). The paper by Kaiser (2021) concluded that deplatforming works because the number of views of channels that were deplatformed by YouTube was lower when they moved to BitChute, a popular YouTube alternative. But this analysis and its conclusion are faulty. The study failed to account for all the other sites that a content creator might go to after being deplatformed. Bitchute is not the only option. Over 2020-21, tens of millions of people abandoned mainstream platforms like Facebook and Twitter and shifted to a growing collection of new offerings that promise less content moderation and more privacy (Spencer, 2021).

Alex Jones, for example, developed several websites and was active on many alternative platforms, including his own mobile app as well as Periscope, Minds, Soundcloud, Scoop, Ello, Thinklink, Bitchute, War Room, Brighteon, TuneIn, Spreaker, Twitch, Radionomy, Bullhorn, and Telegram. After Jones was banned from several big social media platforms, traffic to his Infowars website and app soared (Cobler & Herrera, 2018). However, The New York Times reviewed Infowars's traffic several weeks after the bans and claimed that the intended effects of cutting his reach were working: "In the three weeks before the Aug. 6 bans, Infowars had a daily average of nearly 1.4 million visits to its website and views of videos posted by its main YouTube and Facebook pages. In the three weeks afterward, its audience fell by roughly half, to about 715,000 site visits and video views" (Nicas, 2016).

But the analysis did not include traffic to the two-month-old Infowars app or views of videos that Jones posted on Twitter, where his accounts remained active, and he had largely recovered. In August 2021, 10.6 million individual users visited Infowars (Similar Web, 2021), placing the site's ranking at 1,892 in global internet traffic and engagement and the 412th most frequented site in the United States (Alexa, 2021). He also has several other websites that would enhance those numbers. To properly measure the impact of Jones' deplatforming, therefore, the papers' analysis should have quantified Jones' reach, not only on BitChute after the ban but on other alternative social media sites as well.

There is also a critical flaw in just looking at the number of views Jones receives. The quality of the engagement, for instance, is not shown, but it has a significant impact. For example, if Jones' 10.6 million users in August 2021 went to the website because they aimed to make fun of it or create satire from it, we could conclude that although his engagement is high, people are not engaging positively with it. But if those visitors engage with his content and spread it to others, then he has a positive impact. Put simply, a large number of engagements is indicative of persuasion or reach.

In order to better understand the actual effects of Jones' YouTube deplatforming, data from late July and early August 2021 was reviewed. This helped to better understand the effects of Jones' YouTube deplatforming. The analysis included searching YouTube for Jones, using an in-private browser that was not signed in, and a VPN to mask user data. The top 100 videos returned by the

algorithm were used for data collection and the link, title, and channel hosting the video was recorded, as well as the number of views and the number and ratio of likes and dislikes. Each video was coded as being “for” Jones (either of him speaking or engaged in a civil debate) or “against” Jones (either attacking his beliefs or actions in recorded videos or through parody). The results suggest that, by almost every measure, Jones’s quantitative and qualitative engagement and influence on YouTube in 2021 increased (see table below).

Measures of Engagement for Videos		For	Against	Difference (For – Against)
Mean	View	932712.72	116159.31	816553.41
	Likes	21965.59	3012.03	18953.56
	Dislikes	1164.45	698.57	465.88
	Like-Dislike Ratio	29.46	16.37	13.09
Median	View	227964.00	68720.00	159244.00
	Likes	5400.00	1100.00	4300.00
	Dislikes	149.00	326.00	-177.00
	Like-Dislike Ratio	22.12	7.52	14.60

Median views for videos on the banned Alex Jones Channel in the Bitchute study were fewer than 12,000 with a few outliers receiving more than 1 million views. The analysis documents that in 2021 Jones videos that are “for" him received an average viewership of 932,712, while those against him received an average viewership of 116,159. This means that Jones’ total reach on average for each video is 1,048,871, with the overwhelming majority of those views supporting Jones’ message.

This simple analysis demonstrates the limitations of earlier research that measures the impact of censorship in terms of quantity and fails to consider the quality of engagement. Thankfully, YouTube provides counts of likes and dislikes for most videos, which allows us to judge the extent to which viewers support his speech or those videos against him. Our analysis found that, among

videos that were against Jones, the average number of likes was 3,012 and the average number of dislikes was 698 (with an average like-dislike ratio of 16 — meaning that 16x the number of people "like" the video than "dislike" the video on average). Concerning the videos of Jones himself or videos in support of Jones (coded as "For" rather than "Against"), the videos received an average of 21,965 likes, 1,164 dislikes, and had an average like-dislike ratio of 29—meaning that 29 times the number of people like the video versus disliking the video). If we measure censorship as working by the quality of engagement, not just the quantity, deplatforming is not decreasing Jones' influence. Those videos that are critical of Jones are not performing as well, or even generally well-liked by his fan base, as those that feature Jones' message or in support of it.

The Kaiser (2021) paper also overlooked the cognitive effects of group-think that can occur when controversial content moves from a large platform — where a plurality of viewpoints that push back and critique more extreme perspectives can be shared — to a smaller one. This approach to censorship creates an echo chamber, where individuals grow more certain of their beliefs because they now have limited contact with alternative perspectives, thus further fostering radicalization and extremism (Lane, Shults, and McCaffree, 2021). The burden of banned users swept up on alternative and decentralized online social networks (DOSN) is substantial, yet the blame for extreme content falls largely on the current host, not the past. The social media landscape should not be one where the most powerful and wealthy corporations take no responsibility for providing a platform to the individuals who need the most positive intervention. What would happen if Facebook spent billions of dollars to hire moderators with mental health and deradicalization experience rather than banning such individuals and pushing the problem under the rug?

Research shows that centralized censorship might lead to lower overall engagement but leads to more radical views. One study tracked users of Gab — an American alt-tech social networking service that emerged in 2016. It found that they were more active and hostile after their ban but that their reach declined (Ali et al., 2021). Another study analyzed the data from r/The_Donald and r/Incels, two communities banned from Reddit and subsequently migrated to their standalone websites; moderation measures significantly decreased posting activity on the new platform. However, users in one of the studied communities (r/The_Donald) showed increases in signals associated with toxicity and radicalization, which "justifies concerns that the reduction in activity may come at the expense of a more toxic and radical community" (Ribeiro et al., 2020).

Another study analyzed the consequences of Reddit's banning in 2015 of the subreddits r/fatpeoplehate and r/coontownracist for harassment. The authors found that a proportion of offending users appeared to leave the platform (for Voat, an alternative to Reddit), and that the subreddits that inherited those migrating from those spaces did not see a significant increase in extreme speech. In drawing their conclusions, however, the study's authors stressed that less research had considered the effectiveness of the ban for the health of social media or the Internet at large, as the efforts would make banned users 'someone else's problem,' while pushing them to 'darker corners of the Internet' (Chandrasekharan et al., 2017). Media outlets such as Vice praised the study as proof that censorship works, despite the macro conclusion of the paper, that the speech simply went elsewhere. Of course, speech can be limited in certain forums, but the core question is where does it go, not only what happens to the censored forum?

In cases where there is no opposing feedback to moderate extreme views that come from this renewed feverish interest, echo chambers can breed stronger views with more intense engagement. In other words, the centralized censorship mechanisms which amplify their member's certainty about their beliefs and lessen their openness to alternative views or the people who hold them can combine with the so-called Streisand Effect to further exacerbate tendencies toward extremism.

The Streisand Effect is a fancifully named phenomenon that occurs when efforts to suppress a juicy piece of online information can backfire and end up making things worse for the would-be censor. It comes from a lawsuit initiated by Barbara Streisand against Kenneth Adelman and an organization aiming to raise awareness about coastline erosion, where one of Streisand's seaside homes was photographed on the website (Streisand v. Adelman, 2003). Before her suit, the picture of Streisand's home had been downloaded only six times, two of which are believed to be by her lawyers. By bringing the suit, she brought more attention to the pictures, which were subsequently downloaded tens of thousands of times. Similarly, deplatforming might create greater interest in the person or topic being deplatformed. In psychology this effect is similar to the concept of repression — where unwanted thoughts and unacceptable wishes are repressed by our unconscious and "pushed" out of our awareness. Eventually, these thoughts and wishes find their way out, typically in more inconvenient and harmful manners. Projection can often turn into paranoia, but the point is that repression can result in a counterproductive and uncontained

explosion of the exact thing one was avoiding. Jones seems to have benefitted from similar responses.

What is deplatforming supposed to achieve? If the goal is to curtail the radicalization to violence there is little evidence to support this conclusion. As Jones put it three days after the ban, “The more I’m persecuted, the stronger I get. It backfired” (Nicas, 2018).

1.5 When Censorship Backfires - Echo Chambers, Biases and Victim Mentality

> *"One can radicalize in a manner that is malevolent (e.g. jihadism or neo-Nazism) or benevolent (humanitarianism)." (Reidy, 2018)*

There are several ways that censorship on social media platforms can backfire and promote, rather than suppress, extremism. First, shutting down accounts accused of violating hate-speech policies often shifts those banned individuals to alternative platforms where their grievances are easily aired and more readily accepted, their army of supporters galvanized, and their narrative of long-suffering victimhood further refined. This happened on a mass scale after the 2020 U.S. presidential election and the Capitol riot on January 6, 2021, culminating in the banning of Trump from social media and the termination of cloud services for alternative platforms such as Parler.

Removing radical online content from one platform may decrease the content's reach quantitatively in that platform but exacerbate some of the emotional factors that drive the transition from radical belief to extremist action — such as perceived discrimination, feeling under threat, holding conspiratorial or us-versus-them worldviews, and feeling there is no recourse but violence (National Institute for Justice 2015).

The flashpoint for this violence occurs at an individual level. Research on radicalization — "the processes by which people come to adopt beliefs that not only justify violence but compel it, and how they progress—or not—from thinking to action" (Borum, 2012)— recognizes that the transformation is a "highly individualized process". While radical beliefs do not always precede violent extremist action (McCauley & Moskalenko, 2017), research demonstrates that individual (micro), group (meso), and mass (macro) radicalization share a common bond: They are associated with strong emotional experiences such as anger, shame, guilt, humiliation, fear, love, and hate (Moskalenko & McCauley, 2020). This has important implications for countering violent extremism efforts because it suggests radicalization of opinion and radicalization of action

should be treated separately (McCauley & Moskalenko, 2017). Furthermore, it undermines the deradicalizing value of logical, fact-based refutations of extremist ideologies and recruitment methods. A more fruitful approach might be to recognize the common ground in both empowerment and radicalization (Equal Access International, 2021). One can radicalize in a manner that is malevolent (e.g. jihadism or neo-Nazism) or benevolent (humanitarianism) (Reidy, 2018), with similar emotional mechanisms moving individuals to both legal, nonviolent activism or to extremist violence.

Moral values also affect what is spread on social networks (Pretus et al., 2018). AI algorithms are not currently designed to detect morality, and recent research finds morality is a critical and robust factor in the spread of material online (Brady et al., 2017). This is particularly dangerous for culturally ignorant AI promotion and censorship because moral signatures are important predictors of social, cultural, and political alignment (Haidt & Graham, 2007; Haidt, 2013).

Consider a decision by Tumblr in 2012 that announced the platform would close blogs that promoted self-harm. Proponents of the pro-ana movement — a proactive group of people supporting anorexia, bulimia, and other eating disorders — were forced to find another forum to engage with each other. The result was described by researcher Paola Tubaro: "By forcing blogs to converge into one of the bigger clusters, censorship encourages the formation of densely-knitted, almost impenetrable ana-mia cliques. This favors bonding, but also information redundancy – meaning that pro-ana-mia bloggers will tend to exchange messages, links and images among themselves and to exclude other information sources" (Casilli, Pailler & Tubaro, 2013).

This is a shining example of herd mentality, the tendency of the people in a group to think and behave in ways that conform with others in the group rather than as individuals. Grouping people of dissenting ideas into one closed group reinforces their behaviors and entrenches their thought processes into an endless loop of cognitive dissonance and confirmation bias (Yoo, 2017). In a word, it creates a vibrant echo chamber.

Online networks differ from those in the real world. There are strong and weak ties in any network. Strong ties are characterized as deep affinity — family, friends, or colleagues. Weak ties might be acquaintances or a stranger with a common cultural background or belief. The strength of these

ties can substantially affect interactions, outcomes, and well-being (Granovetter, 1973). Research has shown that negative emotions such as anger are transmitted more efficiently along with weak ties; whereas, joy and happiness are better suited for transmission via strong ties or between those with whom we share real-world experiences (Fan, Xu, & Zhao, 2018). Therefore, negative emotions spread more easily through online social networks. This is not unique to American culture. Research using AI-based sentiment analysis on China's Weibo network also suggests that negative emotions of fear, sadness, and anger are more influential than more cognitively intense discussions on the platform (Song, Dai, & Wang, 2016).

On large-scale platforms, individuals are more likely to have strong ties that reflect their real-world social network (friends and family). That may change once an individual is banned and moves to an alternative, less mainstream outlet, where weaker ties with other users can further facilitate a net negativity bias. Social exclusion creates a painful experience, activating the part of the brain (Dorsal Anterior Cingulate Cortex), where the physical pain is processed (Eisenberger, Lieberman, & Williams, 2003). Individual reflections on painful social experiences (including segregation and exclusion) can lead to increases in psychological variables related to extremism and self-sacrifice, known as "identity fusion" (Jong et al., 2015).

These factors can all converge online. Researchers found that individuals who are susceptible to radicalization lack the ability to distinguish between criticism of their group from criticism of themselves (Lane, Shults, & Wildman, 2018). Criticisms feel personal. This can lead to the adoption of extremist reactions on the part of the offended party (such as self-harming) or a more vitriolic reaction to the opposing values of outgroups (Swann et al., 2012). All of these groups have what are called "sacred values," ideas that they are unwilling to negotiate (Sheikh et al., 2012). Individuals internalize these sacred values as they reflect on personally important experiences, such as feeling excluded or a personal loss, and this reflection causes them to blur the boundary between themselves and their beliefs (Lane, 2021).

Once a group of individuals holds the same sacred values, these ideals become normative markers of group inclusion and identity. When non-radicalized individuals are in a conversation or debate, neither side begins to 1) assume that the other individual's beliefs are paradigmatic of an outgroup or 2) fight for the beliefs with a willingness to be physically harmed. If individuals internalize a belief, those who agree are part of an ingroup, and those who disagree are part of an

outgroup. This can set up a dangerous paradigm where evolutionary mechanisms for group and threat detection and preservation can be triggered by those who hold beliefs that conflict with our own (Shults, 2018; Shults & Gore, 2020). After group boundaries are set, information from an outgroup becomes threatening and even banned — something we have seen in content moderation mechanisms. A study of more than 2 million posts from Twitter and Facebook, for example, documented that posts about outgroups were more likely to be shared across the network and that those that were negatively valenced were more likely to go viral than those that were positive (Rathje, Bavel, & Linden, 2021).

This tracks with research on "affective polarization," the difference between how warmly people view both the political party they favor and the one they oppose. Over the past 40 years, the U.S. has undergone faster polarization than any other Western democracy (Boxwell, Gentzkow & Shapiro, 2020). Use of the internet and social media is often held responsible for these alterations, but, internet usage has risen fastest in countries with diminished polarization. Plus, much of the increase in U.S. polarization is concentrated among older populations with more analog news habits. This suggests that interfering with online engagement across party lines may actually enhance polarization and the us-versus-them worldview that underlies the ingroup/outgroup biases that often drive radicalization (Molenberghs & Louis, 2018). This was exacerbated by the fear and uncertainty that came with the COVID-19 pandemic (Berger, 2020).

Studies show that concerns with evolutionary threats (i.e. predation, contagion, natural disasters, threats to one's social status or economic resources) and anxiety are related to the spread of misinformation and radicalization, as well as nationalism (Bullock et al., 2020). This effect also implicates the development of algorithms and machine learning that is incorporated into policing large-scale social media platforms. Previously flagged posts train AI censorship algorithms (Terdiman, 2018). When coding biases confuse hate speech with speech that is hated, a socio-cultural or political bias is introduced into the system. Studies have shown that political biases and differences across the political spectrum often lead users to flag content, unfriend/unfollow and even distance themselves in the real world (Mitchell et al., 2014; Brown 2020). This is occurring at a time when many social media users, particularly in the U.S., are starting to lessen their belief in the importance of free expression.[5]

[5] A recent global study showed that support of free speech in the U.S. declined more than any other nation except Poland and Hungary https://futurefreespeech.com/interactive-map/. In addition, a study from the PEW Research

In recent years, the policies and practices of the biggest social media platforms are increasingly scrutinized. They have shaped public opinion with experts offering simplistic mono-causal explanations for the relationship between social media usage and radicalization to violence (Gill et al., 2015). These unvalidated assumptions are mostly unchallenged.[6] This suggests that moderation mechanisms have substantial and significant incendiary effects on radicalization online. In addition, radicalization can be inflamed by automated mechanisms that promote the creation and sustenance of isolation and echo chambers. Therefore, greater attention needs to be paid to the automated mechanisms charged with censorship considering the elevated role that deplatforming plays in facilitating online radicalization. AI algorithms can help create polarized echo chambers, which helps explain how radical beliefs emerge from online social network engagement. Of course, AI and machine learning are invaluable tools for facilitating healthy engagements and behaviors as well, even simply informing a user of their cluster, rather than necessarily punishing them for it.

Extremists are a subset of what are referred to as “devoted actors” (Atran & Ginges, 2015). Devoted actors are willing to make extremely costly sacrifices for the cause or group. These sacrifices can be either nonviolent (e.g. getting arrested, beaten by police, losing one's job) or violent (e.g. carrying out a terrorist attack, violently protesting, targeted assassinations). Devoted actors include followers of Gandhi, Mandela, Martin Luther King Jr., as much as followers of ISIS or neo-Nazis. Many people who join violent groups follow a trajectory whereby they could have joined a nonviolent devoted actor group (Gomez et al., 2021). The key for respecting civil liberties while mitigating political violence is knowing how to nudge those towards non]violent, or ideally, pro-social devoted actor groups. Research on the psychology of devoted actors shows that they have a combination of "sacred values" (i.e. moral values of the highest importance) and "fused" identities (i.e. a visceral personal connection) with a small band of friends (Gomez et al., 2017; Atran et al., 2015; Sheikh et al., 2016).

Forum found that Millennials are more likely to support limits to free speech than older generations: https://www.pewresearch.org/fact-tank/2015/11/20/40-of-millennials-ok-with-limiting-speech-offensive-to-minorities

[6] As Audrey Alexander, researcher and instructor at the Combating Terrorism Center in West Point stated in a webinar at George Washington University’s Program on Extremism entitled, Global Internet Forum to Counter Terrorism: Balancing Online Content Moderation and the Rule of Law, “If we start providing policy problems to what we believe is the problem, instead of what is actually the problem, that’s when we get in over our head and start developing solutions that we don’t really need and that don’t really work…really in the literature there is not a compelling relationship where terrorist content causes activity.” https://www.youtube.com/watch?v=ScOmMjEB-yY

Brain scan research done on jihadist supporters showed that feelings of social exclusion can increase the chances of becoming devoted actors (Pretus et al., 2018). But altering their perception of social norms (i.e., what they think their peers think) reactivated the areas of the brain associated with deliberation and self-reflection, and caused them to lower their explicit willingness to fight and die for these sacred values (Hamid et al., 2019).

This demonstrates that excluding extremists, especially those at an early stage, from public discourse platforms can backfire and increase propensity towards violence, both neurally and behaviorally. Conversely, shaping their perceptions of what they think their peers think can actually reduce their violent intentions. The key takeaway from this research is that the social norm interventions need to be directed at potential actions rather than the values themselves. Research in Israel and Palestine found that attempting to negotiate with sacred values using material incentives can backfire by increasing moral outrage and willingness to use violence against the other side. Conversely, offering symbolic concessions decreased the outrage and increased willingness to negotiate and speak with the opposing side to find a compromise to the conflict (Ginges et al., 2007). The neuroscience research builds on these findings. It shows that trying to change the values of extremists may not be necessary to reduce violence and that offering an alternative action pathway may be sufficient for finding a peaceful solution (Atran 2020).

What does this mean for social media companies? Don't exclude those with perceived radical views — engage with them whenever possible prior to violent threats. One way is to use entertainment with social norms embedded in them, as has been demonstrated in successful media interventions in Rwanda (Paluck 2009), the DRC (Paluck 2010), and Nepal (Clark et al., 2018; Cislaghi et al., 2019). Another option is to amplify the voices within an ingroup that denounce violence (Welzel and Deutsch, 2012). There are many intervention techniques that have been successfully implemented in other contexts — like trying to reduce smoking, alcohol abuse, or delinquency — that can be adjusted for implementation into the countering-extremism space (Tankard and Paluck, 2017).

In the paper "Neural correlates of maintaining one's political beliefs in the face of counterevidence" by Kaplan, Gimbel and Harris (2016), neural imaging was used to investigate the neural systems involved in maintaining beliefs in the face of counter evidence. The conclusions

showed that individuals with firm beliefs can become more attached to those beliefs when confronted with contradictory information. This opens up additional complexity with regard to the process of deradicalization. It would be worth considering the evolution of the neurological response after encountering repeated counter evidence over a long period of time.

Finally, in the Vox piece “Does banning extremists online work?” (Ghaffary 2022) the author provides a generally balanced array of evidence of both how specific influencer reach can be limited and also how the censored users can become more radicalized. The article states, “Several academic studies in the past few years have also quantitatively measured the impact of major social media networks like Twitter, Reddit, and YouTube deplatforming accounts for posting violent, hateful, or abusive content. Some of these studies have found that deplatforming was effective as a short-term solution in reducing the reach and influence of offensive accounts, though some studies found increases in toxic behavior these users exhibited on alternative platforms.”

1.6 Radicalization and Violent Extremism Cannot be Treated as One

In conducting this multidisciplinary review on radicalization in social media, the focus was on the academic and research community's longstanding work on radicalization, extremism, and political violence.

But a critical approach to radicalization and extremism studies is sorely lacking and markedly inconsistent. Thus, influential organizations like the Anti-Defamation League (ADL) can run a massive campaign against Facebook to Stop Hate for Profit, while simultaneously publishing a scathing piece claiming that BitChute, a video sharing platform established in reaction to social media censorship, has become a "hotbed for violent, conspiratorial and hate-filled video propaganda, and a recruiting ground for extremists" (Anti-Defamation League, 2020). This echoes the research sentiment that, "policing within a single platform (such as Facebook) can make matters worse, and will eventually generate global 'dark pools' in which online hate will flourish" (Johnson et al., 2019).

Current methods of social media censorship worsen the problem. While these social media companies claim their methods derive from an extremely large body of research, much of what they tag as removable content that has violated their standards has a reasonable assumption of political partisan alignment.

It is essential to differentiate between cognitively radical and violent extremist expressions. It is modeled on the differentiation under American law between legal hate speech and illegal threatening communications, one which appreciates complexity and foresees the second and third-order effects of prohibition.

Governmental, academic, and activist communities have long tried to identify factors that motivate people to embrace radical views and act upon them violently. But the results of these

varied inquiries have proven unempirical. A review of the literature on radicalization, extremism, and political violence reveals that in the field of extremism studies, radicalization is a highly complex, non-linear and individualized process (Horgan, 2008).

A recent systematic review of thousands of papers revealed that "empirical evidence of variables that discriminate between terrorists and non-terrorists is limited," and a focus on the risk factors identified in the study "might contribute to discrimination and reduce the effectiveness of counterterrorism strategies."

Despite the absence of reliable supporting evidence, many conceptual models for radicalization exist. Few of them, however, seem to be derived from empirical research (Borum, 2012). In addition, almost all were developed in a period when the only true concern was jihadist terrorism. Yet the same approach and models (with accompanying mistakes) are being applied to root out far-right, far-left, and ethno-supremacist extremism.

Before the War on Terror was declared in the aftermath of the horrendous terrorist attacks of September 11, 2001, there were few references to the word "radicalization" in academic or journalistic literature. An obsession to determine the roots of radicalization after 9/11, alongside the occupation of Afghanistan and Iraq that followed, gave Osama bin Laden everything he wanted. Terrorism is theater. As terrorism researcher Brian Jenkins put it, "Terrorists don't want a lot of people dead; they want a lot of people watching" (Violent Islamist Extremism, 2007). Today, that mindset which launched the War on Terror is used to understand domestic (as opposed to foreign) extremism (Mueller & Stewart, 2018). Radicalization is still a burgeoning area of research but there has been a decreased focus on jihadist extremism in recent years. The election and defeat of Donald Trump, an increase in acts of far-right (and left) wing terrorism, and other variables have caused many to try to understand radicalization as it relates to domestic groups and ideologies. If the same "strategic blunders" (Bumiller, 2008) that marked the post-War on Terror era are replicated in the fight againste domestic extremism, matters could get even worse. As mentioned in the introduction, there are nearly four times as many Sunni Islamic militants today as there were on September 11, 2001 (Jones et al., 2018). That might be accelerated if hate speech or extremist social media content is aggressively censored. (Reed & Ingram, 2018). Radicalization is extremely complex, and the role social media plays in radicalization is only part of the equation.

As one expert on the psychology of terrorism explained: “In our preoccupation with radicalization, we still know far less about aspects of the terrorist 'arc' of involvement than we should by now. We still know relatively little on the specific psychological and social dynamics that propel individuals to take action on behalf of such groups (and what sets them aside from those who seek to remain involved, but not in this particular way). In essence, it is ironic that we focus on those issues that are more resistant to actual behavioral change, and thus, we ignore those issues that are far less resistant to imaginative interventions” (Borum, 2011).

These complexities make it tricky to define and find solutions to the problem of radicalization (Neumann, 2003). Simply put, “radicalization" describes a socialization process that sometimes leads to violent extremism. But countering extremism does not counter terrorism — it counters the process of ideological adoption. Yet violent extremist action remains relatively low, and there are few actual terrorists relative to the number of those radicalized (Taylor, 2010; Taylor & Currie, 2012). Only a small number of “cognitively radicalized” individuals become destructive terrorists. We pay attention to the few violent extremists (Davis et al., 2013) rather than investigating how or why the vast majority of others did not become (violent) extremists (Reidy, 2019; Bosley, 2020).

Investigating the violent extremist may seem logical but this mindset exaggerates the likelihood of “radicalized” individuals engaging in terrorist activity (Horgan, 2012) and may be counter-productive (Sageman, 2015). For example, countering violent extremism (CVE) initiatives have stigmatized Muslims (Brennan Center for Justice 2015), thus accelerating both radicalization and violent extremism (Sjøen & Jore, 2019). This solidifies a victim mentality that enhances ingroup/outgroup bias (Whitbourne, 2010), further dehumanizing outgroup members (Kteily, Hodson, & Bruneau, 2016) and justifying the turn to violence for potential extremist actors (Varvin, 2005). Conflating cognitive radicalization and violent extremism can backfire and lead to a diverse and largely nonviolent group being labeled as “violent extremist.” Enhanced online scrutiny may in turn inflame radicals.

In science, it is crucial to understand that correlation does not equate to causality. While many violent extremists or terrorists radicalized, at least in part, online and often engage with radical or violent extremist content over social media, there is a difference between radicalizing while using social media and being radicalized by social media. Always remain skeptical of simplistic

mono-causal explanations. The fact that extremists use the internet is indicative of modern life; it is not unique to extremism. In 2018, for example, 90% of American adults aged 18-29 used social media (Smith & Anderson, 2018). One might also ask whether ever-expanding social media censorship or perceptions of discriminatory social media company gatekeeping correlate to a heightened interest or susceptibility to extremist content or conspiracy theory. Unfortunately, there seems to be no interest or funding for such inquiries. When such research is conducted, such as the internal study from Facebook that revealed that its algorithms were promoting division and echo chambers on the platform, they have been shut down and unsupported (Horwitz & Seetharaman, 2020).

Free expression, and even having a society with stable balanced ideas that can be rationally debated, is threatened by social media censorship. Therefore, social media censorship, and not social media itself, may represent a threat to democracy. Social media companies are private corporations, so they have a right to determine what is hosted on their platforms. But when that censoring is applied disproportionality or equitably it is more akin to social engineering. Furthermore, when social media companies' policies are based on flimsy evidence for the effects of censorship or ignorance of scientific evidence that suggests policing might make problems worse, it feels like the right time to build an alternative approach and to test it.

An extremely low rate of those cognitively radicalized go on to commit acts of violent extremism. Eradicating "radical" content may reduce the risk of violent extremist action, but censoring speech only makes communities feel threatened and under attack, and states of fear positively correlate to the commission of violent extremism (Pauwels & Heylen, 2020). If policing on mainstream social media platforms might worsen matters, over the long-term especially, then suppressing an individuals' ability to express radical or hateful ideas online might exacerbate violent extremist behavior.

Instead, allowing radical content that does not constitute a threat under U.S. law or our interpretations of the First Amendment might improve a preventative factor in radicalization to violence. Thus, an alternative model for social media platform content moderation that differentiates between radical and violent extremist content is due. Rather than just allocating resources to human moderators and relying on AI and machine learning, we propose directing resources and efforts to combat hate with alternative messaging in order to provide an off-ramp

to radicalization. Such an approach can be more effective than current censorship and gatekeeping.

This is what the George Washington University study (Johnson et al., 2016) indicates: "when attacked, the online hate ecology can quickly adapt and self-repair at the micro-level." The researchers observed that, "operationally independent platforms — that are also commercial competitors — [can become] unwittingly coupled through dynamical, self-organized adaptations of the global hate-cluster networks." The researchers witnessed how banned Facebook users migrated to less-policed platforms and reentered the platform through "back doors,' consistent with the adaptations made by the Boogaloo Boys, ISIS, QAnon, and other groupings. As the authors noted, "the existence of several operationally independent platforms with their own moderators and no coordinated cross-platform policing gives rise to a further resilience," and, "shows that sections of the less policed platforms can then become isolated, creating spontaneous 'dark pools' of hate highways."

Previous research by the same authors documented the resilience of ISIS supporters in 2015-2016, an online collective that remains resilient and active despite massive pressure to suppress it (Johnson et al., 2016). The internet will always provide places for extremists to organize and engage.

They suggest a "matrix of interventions according to the preferred top-down versus bottom-up approach on a given platform and the legal context in a given country." Proposed interventions include the encouragement of, "anti-hate users to form clusters…which then engage in narrative debate with online hate clusters. Online hate-cluster narratives can then be neutralized with the number of anti-hate users determined by the desired time to neutralization" (Johnson et al., 2019). Their modeling of this policy suggested a decrease in hate speech over time. We also believe that such an approach based on anti-hate dialogue and engagement would not only mitigate the size of online hate clusters, but drastically decrease the number of radicalized individuals engaging with violent extremist content online who would go on to mobilize for, or commit real-world acts of extremist violence.

The strategy of Daryl Davis — who deradicalized hundreds of KKK members by befriending them, listening and engaging in respectful dialogue — is essential to understanding how a long-term

approach can work, both digitally and physically. There cannot be expectation of immediate change in ideology. Minds change over long periods of time with steady information exchange. Davis engages in both physical and digital communication strategies with his contacts and both have proven to assist.

Current trends in social media gatekeeping create a technocratic top-down approach to subjectively interpreting what is and what is not “hateful” content. Rather than reducing hate and enhancing social cohesion, such praxes may worsen matters.

The Change Minds Initiative moderation model distinguishes between radical and violent extremist content in ways that parallel the differentiation between free speech and threatening communications under U.S. law. It also draws energy from the belief that free expression remains an inherent natural right. Democracies depend on it. As such, the Minds model focuses on anti-hate dialogue, critical thinking training and engagement as a means of reducing the risks of radicalization on our platform. As Martin Luther King Jr. put it, “Returning hate for hate multiplies hate, adding deeper darkness to a night already devoid of stars. Darkness cannot drive out darkness; only light can do that. Hate cannot drive out hate. Only love can do that” (King Jr., 2019).

1.7 A New Approach to Moderation

Society can't just shut off the lights and hope violent extremism goes away — it won't. Censorship is avoidance, and avoidance never solves a problem. It may even make it worse.
The move to violent extremism can never be solved with simplistic answers to complex scenarios. There must not be monocausal determinations of the link between social media usage and radicalization. And a single platform analysis won't do. Our world doesn't need a bigger "Whack-a-Mole" mallet nor to cover the mole holes — the mole will eventually find another hole through and cause even more havoc.

Every social network's content moderation practice is failing. Some over-moderate, and others essentially do far less than they should. Some employ unreliable algorithms that mistakenly take down innocent content. Large social networks have abused their power with inconsistent and unreliable standards. Current approaches are likely to make matters worse by cementing grievances, and heightening insecurity, uncertainty and perceptions of threat among the most vulnerable. Censorship pushes extremists deeper into echo chambers, promoting isolation, frustration, and motivation by bad actors.

Our world needs an alternative method for content moderation that can be tested, refined and documented with an appreciation for the complexities of radicalization into violence.

The Change Minds Initiative framework is guided by several propositions:

1. Radical agents interact primarily with other radical agents when banned from interacting with neutral agents;
2. New radical users on any social network site become more radical when forced into an echo chamber (defined broadly by their subscription/subscriber centricity);
3. When radical users are merged with more users, their sentiment and subscriber/subscription centricity become less radicalized;

4. Radical user sign-up on Minds is sometimes driven by their banning from other social media sites, and this results in a potential increase in risk factors associated with radicalization to violence;

5. Distinguishing between cognitive radicalization (extreme beliefs) and violent extremism (clear threat, support or engagement) is key to the development of effective alternatives to content removal and platform policing, and

6. Hate speech can be harmful, but a whole-of-community approach offers a better means of addressing the underlying psychological and socio-ecological factors than censorship (particularly in the long-term).

Once clear metrics distinguish between cognitively radical and violent extremist material are delineated, AI and other technological tools can contribute to the solution. However, creating whole-of-society, peer-to-peer humanist and holistic solutions should be a priority, particularly if the short and long-term are considered.

Minds has established a framework that encapsulates our conclusions. This framework explores the practical side of creating a new methodology for moderation and combating hyperpolarization, hate, and violent extremism online.

1.8 The Minds Moderation Framework

Minds hopes to elevate global discourse through Internet freedom. A fair moderation system must first prioritize trust between its users and the platform, creating a foundation on which people can build steady connections and mutual understanding. The key to establishing this trust is a transparent and predictable moderation framework that is based on the core principles of Internet freedom and draws upon academic data.

Open Source

Open source software is at the core of providing a fair moderation system. Minds is licensed under AGPLv3, a free, copyleft license published by the Free Software Foundation. All of its source code is publicly available at developers.minds.com where anyone is free to contribute to the project, use the code in other applications, or inspect the code to understand how things work. Gitlab is where Minds conducts all of its software development and project management, in the open for public scrutiny and participation.

Allowing people outside the company to view the code is a significant step towards more trust and accountability. Open source enables anyone to inspect the algorithms that power Minds to ensure the app is delivering what users expect and not maliciously tracking and targeting. It also enables anyone to contribute to the algorithms, breeding more collaboration and innovation between the network and its community.

Build Your Algorithm

The algorithms that govern what you see on social media platforms like Facebook and Twitter are centrally-controlled and non-transparent. While these algorithms serve the purposes of keeping feeds fresh and making it easier to discover new content and creators, they come with the costs of a lack of control on the user's end and the resulting suspicion that our social media -- our windows into reality -- are influenced and shaped by biases and decisions we don't agree with.

Minds puts you in control of your own algorithm. By setting your content preferences, you decide what types of content do and don't get amplified for you. And by answering a few questions about your personal outlook on things like censorship, misinformation, government regulation, and optimism for the future, it enables Minds to better connect you to content and creators you haven't discovered yet and that may become your new favorite voices on Minds.

A key question in the Build Your Algorithm tool is how open you are to seeing opinions that you disagree with. This is the first time any social network has provided a preference such as this, which intends to allow users to break free from their echo chambers and challenge their own opinions more deeply. When someone answers this question, they are required to either acknowledge their own closed-mindedness or open their algorithm up to more diverse opinions. These are crucial steps towards encouraging dialogue and elevating global discourse on social media.

In the long term, this data will be essential in effectively measuring open-mindedness and user preference for dialogue or breaking out of echo chambers. But ultimately the tool is about providing the user with control and transparency over their own algorithm, rather than relying on a middleman to decide on these preferences for you without any accountability.

Free Expression

The First Amendment is one of the most essential and battle-tested content policies in human history. Freedom of speech is at the foundation of American democracy and there has been over a century of legal casework built on top of it.

Minds bases its content policy on the First Amendment to ensure that information is only removed when breaking the law or attempting to maliciously hack/attack the app. This can broadly be categorized as content that promotes terrorism, paedophilia, extortion, fraud, harassment, revenge porn, sex trafficking or imminent acts of violence. Other categories include copyright infringement, trademark infringement, impersonation spam and malware. Any Minds user can report a piece of content for falling under these categories at any time. In the event that

content is removed for one of these reasons, further action may be necessary such as contacting local authorities, depending on the individual case.

Other cases where content removal may be required depending on the individual case include more nuanced categories such as spam or inauthentic behavior, such as using fake engagement, bots or scripts to abuse the system. Minds is continuing to research and develop new ways to handle these cases that are fair and transparent and to provide the community with a voice in the decision making process.

Minds users are empowered with robust filtering and blocking tools to curate their feeds how they desire. Tools like user blocking and reporting are essential to empower users to protect themselves or flag content that may break the terms of service. NSFW (Not Safe For Work) filters and controls provide more user control and categorization of controversial content. It is worth noting that not all Big Tech apps are unified in their censorship approaches. While they do mostly align on hate speech and misinformation policies, Twitter for instance does allow pornography.

Protecting freedom of speech for social media users is essential to create a fair playing field and a breeding ground for change. It ensures that there is no subjective middleman interfering with access to information, and it creates the opportunity for real dialogue to occur. Free speech means radical, extremist content is also protected. Censoring this content has many adverse effects and the evidence in this paper clearly demonstrates how it can result in more radicalization.

Allowing extremist content to exist creates an identifiable entry point for a positive intervention or the start of a dialogue. It creates an actual opportunity for change, whereas censorship eliminates opportunity. We must face our societal problems head on, and ensure that social media platforms are providing this opportunity for change and not preventing it.

Privacy

Privacy is foundational to provide users with control. Minds allows users to create pseudonymous accounts if they wish, along with the option using a real identity. Additionally, Minds structures its data pipeline to leverage pseudonymous user identification numbers to ensure that any digital

profile can not be matched with a public username on the Minds network. This is in stark contrast to the invasive surveillance techniques leveraged by large technology companies to build psychographic profiles and sell access to that data to advertisers and others.

Pseudonymity allows users to separate their digital and physical identities, which can help them feel more safe and comfortable to engage with others. If we accept that free speech enables positive change, and we recognize that speaking freely often requires that people feel in control of their privacy, then we must also secure privacy to secure free expression.

Minds also provides users with a messaging system, Minds Chat, which is built on the Matrix protocol with end-to-end encryption, meaning only the end user has the key to access the content of the messages. This level of privacy protects users from having their private conversations leaked, and it also creates an environment for more direct conversation between two or more individuals. It also ensures that the company or platform itself does not have access to the content of the direct messages. Encrypted messaging is an important element to provide users with a safe place to have difficult conversations, and an effective tool for positive intervention with users trying to cope with mental illness or who are at risk of radicalization to violence.

Community Governance

In networks where centralized administrators can ban users or remove content, it is crucial to have a fair and transparent process. Mistakes may happen, or worse, centralized abuse can develop, and a transparent process with community voice is an effective check against these risks. Currently, Minds leverages a community jury system to review appeals on moderation decisions. This enables the community to essentially reverse a decision made by Minds moderators should the jury reach a consensus that the initial action was inconsistent with the network's policy.

The jury system creates an essential checks-and-balances system on the entity responsible for moderation and ensures that all users have a voice in the event of an appeal. It also ensures that the appeal is reviewed by a different entity than the one which made the initial decision.

IMinds aims to expand on its community governance system and provide a method for community members to establish trust, which in turn can grant them more decision making power across various aspects of the network. These decisions include areas such as the removal of content, the tagging of content, the paid promotion of content, the ranking of content and more.

Crypto Incentives and the Decentralized Future

Distributed systems and blockchain technologies such as Bitcoin and Ethereum have effectively eliminated the middleman from payments, providing the user with ultimate control over their transactions, financial and other. The transactions are all publicly auditable on the blockchain to provide maximum transparency. They also provide networks with a new mechanism to easily share value with its contributors.

Additionally, the emergence of immutable databases makes censorship even more of a losing battle. Minds currently harnesses Ethereum to provide MINDS token incentives for activities such as jury participation, reports, curation and moderation. It is unsustainable to expect that a community-driven project can keep up with the giants of tech without rewarding community participants.

Building a gamified token reward system is a very effective method to incentivize or discourage certain behavior. This conceptually is a very important piece of the puzzle in creating desired outcomes, such as healthier discourse and less polarization. It enables the network to truly share its value with the community, and for the community to be fairly rewarded for the time and energy they put into it. We're early in the development of these systems, and see token rewards as an alternative incentive to attention, which we associate with the bad outcomes of polarized discourse.

Over time, both the governance logic and infrastructure of Minds will be migrating to more resilient peer-to-peer systems, but both centralized and decentralized technologies are essential to building an enterprise-grade and scalable experience.

Self-Sovereignty

Users of the network should ultimately have control of their identity, content and social graph. If this can be achieved, then users are no longer “locked-in” with any individual network and can choose where to bring their value without fear of losing anything. This forces social networking platforms to compete for the trust of a shared user base, making it a more fair playing field.

Self-sovereignty enables ownership and allows for value to be accrued by the individual rather than the platform. When users do not have vendor lock-in, they are in control of the value they provide. Self-sovereignty, along with community governance, are areas that Minds is deeply researching, with the ultimate goal of providing portability for its users and for the network itself. It will likely take more experimentation and time for these technologies to mature before they can be implemented at scale. Minds plans to continue to invest time and energy into the development of this framework which includes standards like decentralized identity (DID) and verifiable credentials (VCs).

Events

MINDS IRL is an ongoing event series to generate constructive dialogue across the political spectrum and bring digital debates to the physical world. The first event was titled "MINDS IRL: Ending Racism, Violence and Authoritarianism". It was headlined by Daryl Davis and brought together dozens of influential Internet voices for panel discussions and breakout sessions.

The event was initially deplatformed by the Pitman Theater in New Jersey due to protests from Antifa, where the group went so far as to call Daryl Davis a ‘white supremacist’ which he responded to in a post on his Minds channel. Ultimately the event moved to Philadelphia, where it was sold out and generated significantly more media attention than it would have without the deplatforming. Live events have been disrupted by the COVID-19 pandemic, but will continue as early as possible.

Partnerships

Joining forces with other networks, individuals and organizations is the only possible way this type of holistic approach, cross-internet, can work. Consistency across platforms brings down the global temperature.

So far, Minds has collaborated with Tech Against Terrorism, Parallel Networks, Braver Angels, Light Upon Light, Daryl Davis, Tasman AI, and CulturePulse, and invites anyone with interest in participation to contact info@minds.com.

Calls to Action

- Participate in the #ChangeMinds by creating a post on social media that tells the world how you changed your mind about anything from a major global issue to your favorite music. See more information at www.minds.com/change.
- Encourage other social media platforms to adopt this, or a similar, moderation framework in order for cross-network impact to occur, and to prevent controversial content from simply being pushed from one social network to another.
- Report any errors or proposed improvements to this paper to info@minds.com.

1.9 Contact

- https://minds.com
- https://support.minds.com
- https://developers.minds.com
- https://minds.com/change
- https://twitter.com/minds
- https://www.culturepulse.ai

1.10 Prospects for Future Research

- The correlation between isolation, both digital and physical, with violent behavior
- Further evidence of direct correlation between censorship and violent extremism
- How to prevent social media addiction
- Understanding the variation in effectiveness of deplatforming to limit the reach of an individual or group depending on their audience size and interrelated events which impact the Streisand effect
- The relationship between online and offline radicalization and social networks
- What is positive radicalization?
- Memetic evolution of symbols and language after censorship
- The impact of providing mental health resources to users on Minds and other networks
- The line between legal and illegal of various speech in the USA and abroad
- The impact of community jury systems on community sentiment
- Analyzing the rate of deradicalization over time on Minds compared to restrictive apps
- Do people change their minds more easily in a free speech environment?

Appendix: How Large Social Media Platforms Define Hate Speech

Twitter's definition of hate speech:

> *"You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.*
>
> *Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category." (Twitter, 2021)*

Facebook's definition is:

> *"We define hate speech as a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we*

provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic.

We recognize that people sometimes share content that includes someone else's hate speech to condemn it or raise awareness. In other cases, speech that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content." (Facebook, 2021)

Google defines hate speech as:

"Content that promotes hatred, violence, & attacks against groups of race or ethnicity, age, religious beliefs, disability, gender, sexual orientation, gender identity, or veteran status is considered hate speech." (Google, 2021)

However, its subsidiary YouTube, which is more engaged in content moderation of their users, has a more lengthy definition of hate speech, which also is expanded to include other forms of censorable content:

"Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes:

- *Age*
- *Caste*
- *Disability*
- *Ethnicity*
- *Gender Identity and Expression*
- *Nationality*
- *Race*
- *Immigration Status*
- *Religion*
- *Sex/Gender*
- *Sexual Orientation*

- *Victims of a major violent event and their kin*
- *Veteran Status*

If you find content that violates this policy, report it. Instructions for reporting violations of our Community Guidelines are available here. If you've found a few videos or comments that you would like to report, you can report the channel.

What this policy means for you. If you're posting content - Don't post content on YouTube if the purpose of that content is to do one or more of the following:

- *Encourage violence against individuals or groups based on any of the attributes noted above. We don't allow threats on YouTube, and we treat implied calls for violence as real threats. You can learn more about our policies on threats and harassment.*
- *Incite hatred against individuals or groups based on any of the attributes noted above.*

Other types of content that violate this policy include the following:

- *Dehumanizing individuals or groups by calling them subhuman, comparing them to animals, insects, pests, disease, or any other non-human entity.*
- *Praise or glorify violence against individuals or groups based on the attributes noted above.*
- *Use of racial, religious or other slurs and stereotypes that incite or promote hatred based on any of the attributes noted above. This can take the form of speech, text, or imagery promoting these stereotypes or treating them as factual.*
- *Claim that individuals or groups are physically or mentally inferior, deficient, or diseased based on any of the attributes noted above. This includes statements that one group is less than another, calling them less intelligent, less capable, or damaged.*

- *Allege the superiority of a group over those with any of the attributes noted above to justify violence, discrimination, segregation, or exclusion.*
- *Conspiracy theories saying individuals or groups are evil, corrupt, or malicious based on any of the attributes noted above.*
- *Call for the subjugation or domination over individuals or groups based on any of the attributes noted above.*
- *Deny that a well-documented, violent event took place.*
- *Attacks on a person's emotional, romantic and/or sexual attraction to another person.*
- *Content containing hateful supremacist propaganda including the recruitment of new members or requests for financial support for their ideology.*
- *Music videos promoting hateful supremacism in the lyrics, metadata, or imagery.*

(YouTube, 2021)

References

Aiello, E., Puigvert, L., & Schubert, T. (2018). Preventing violent radicalization of youth through dialogic evidence-based policies. International Sociology, 33(4), 435–453. https://doi.org/10.1177/0268580918775882

Alexa. (2021) "Infowars.com." Alexa, https://www.alexa.com/siteinfo/infowars.com

Alexander, A. & Braniff, W. (2018, January 21). "Marginalizing Extremism Online." Lawfare Blog, Foreign Policy Essay. https://www.lawfareblog.com/marginalizing-violent-extremism-online

Ali, S., Saeed, M.H., Aldreabi, E., Blackburn, J., Cristofaro, E.D., Zannettou, S., & Stringhini, G. (2021). Understanding the Effect of Deplatforming on Social Networks. 13th ACM Web Science Conference 2021.

Anti-Defamation League (2020, August 31). "BitChute: A Hotbed of Hate," ADL Blog, Anti-Defamation League. https://www.adl.org/blog/bitchute-a-hotbed-of-hate.

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. Communication Monographs, 77(3), 281–311.

Bauerlein, M., & Jeffery, C. (March -April, 2021). "Why Facebook Won't Stop Pushing Propaganda, " Mother Jones: Politics. https://www.motherjones.com/politics/2021/08/why-facebook-wont-stop-pushing-propaganda/?scrolla=5eb6d68b7fedc32c19ef33b4

Baugut, P., & Neumann, K. (2020). Describing Perceptions of Media Influence among Radicalized Individuals: The Case of Jihadists and Non-Violent Islamists. Political Communication 37(1): 65–87. DOI: 10.1080/10584609.2019.1663323

BBC. (2021). "Facebook Bans ' Voice of Trump' from Platform." BBC News: US & Canada. https://www.bbc.com/news/world-us-canada-56598862.

Benkler, Y., Faris, R., & Roberts, H. (2018). Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press. ISBN-10 : 0190923636

Berger, JM., & Morgan, J. (2015). The ISIS Twitter Census: Describing the Population of ISIS Supporters on Twitter. Brookings Institute - Center for Middle East Policy (Analysis Paper). https://www.brookings.edu/research/the-isis-twitter-census-defining-and-describing-the-population-of-isis-supporters-on-twitter/

Berger, JM., & Perez, H. (2016). The Islamic State's Diminishing Returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters. George Washington University - Program on Extremism (Occasional Paper).

Berger, JM. (2020, October 9). "Our Consensus Reality has Shattered." The Atlantic, https://www.theatlantic.com/ideas/archive/2020/10/year-living-uncertainly/616648/

Bolsen, T., & Druckman, J. N. (2015). Counteracting the Politicization of Science. Journal of Communication, 65(5), 745–769. https://doi.org/10.1111/jcom.12171

Bond, S. (2021, May 5). "Facebook Ban On Donald Trump Will Hold, Social Network's Oversight Board Rules," NPR: Technology. https://www.npr.org/2021/05/05/987679590/facebook-justified-in-banning-donald-trump-social-medias-oversight-board-rules

Borum, R. (2012). Radicalization into Violent Extremism I: A Review of Social Science Theories. Journal of Strategic Security 4(4): 7–36. https://doi.org/10.5038/1944-0472.4.4.1.

Bosley, C. (2020). Violent Extremist Disengagement and Reconciliation. United States Institute of Peace.

Boxwell, L., Gentzkow, M., Shapiro, J.M. (June 2020; August 2021, rev.) Cross-Country Trends in Affective Polarization. National Bureau of Economic Research, Working Paper 26669, DOI: 10.3386/w26669

Brady, W. J., Wills J.A., J. T. Jost J.T., Tucker, J. A. & Van Bavel J. J.(2017). Emotion Shapes the Diffusion of Moralized Content in Social Networks. Proceedings of the National Academy of Sciences 114(28): 7313–18. https://doi.org/10.1073/pnas.1618923114.

Brennan Center for Justice.(2015). “Countering Violent Extremism (CVE): A Resource Page.” Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/countering-violent-extremism-cve-resource-page.

Brennan, M. (2021, February 18). “Views of Big Tech Worsen; Public Wants More Regulation,” Gallup: News. https://news.gallup.com/poll/329666/views-big-tech-worsen-public-wants-regulation.aspx?utm_source=alert&utm_medium=email&utm_content=morelink&utm_campaign=syndication

Brown, A. (2020, April 24). “Most Democrats who are looking for a relationship would not consider dating a Trump voter.” Pew Research Centre, https://www.pewresearch.org/fact-tank/2020/04/24/most-democrats-who-are-looking-for-a-relationship-would-not-consider-dating-a-trump-voter/

Bruneau, E., Kteily, N., & Falk, E. (2018). Interventions highlighting hypocrisy reduce collective blame of Muslims for individual acts of violence and assuage anti-Muslim hostility. Personality and Social Psychology Bulletin, 44(3), 430–448.

Buckley, T. (2015). “Why Do Some People Believe in Conspiracy Theories?” Scientific American. 10, 1 July.

Bullock, J., Lane, J.E. Mikloušić, I., & LeRon Shults, F. 2020. “Modelling Threat Causation for Religiosity and Nationalism in Europe.” ArXiv:2009.09425 [Physics], September. http://arxiv.org/abs/2009.09425.

Bumiller, Elisabeth. (April 9, 2008). “At Hearings, a Chance to Explain Iraq Views and Audition as Commander in Chief.” The New York Times: Washington. https://www.nytimes.com/2008/04/09/washington/09scene.html.

Carlson, CR., & Rousselle, H. (2020) Report and repeat: Investigating Facebook’s hate speech removal process. First Monday 25(2 - 3) doi: http://dx.doi.org/10.5210/fm.v25i2.10288

Casilli, AA., Pailler, F.; Tubaro, P. (2013) Online networks of eating-disorder websites: why censoring pro-ana might be a bad idea. Perspectives in public health, 133(2), 94–95. https://doi.org/10.1177/1757913913475756

CBS News. (2019, June 6). “English-speaking ISIS supporters exploit messaging app.” CBS News. https://www.cbsnews.com/news/english-speaking-isis-supporters-exploit-messaging-app/

Chappell, B., & Tsioulcas, A. (2018, August 6). “YouTube, Apple and Facebook Ban Infowars, Which Decries 'Mega Purge',” NPR: Media. https://www.npr.org/2018/08/06/636030043/youtube-apple-and-facebook-ban-infowars-which-decries-mega-purge

Chandrasekharan, E, Pavalanathan, U, Srinivasan, A, et al. (2017) You can’t stay here: The efficacy of Reddit’s 2015 Ban examined through hate speech. Proceedings of the ACM on Human-Computer Interaction 1(2): art. 31.

Chibelusi, W. (2021, September 5). “The Taliban has over 800k Twitter followers - could they get a blue tick soon?” ITV News: World. https://www.itv.com/news/2021-09-04/the-taliban-has-over-800k-twitter-followers-could-they-get-a-blue-tick-soon

Clarke, L., & Swindells, K. (2021, June 9). “How social media companies help authoritarian governments censor the internet,” The New Statesman: Business. https://www.newstatesman.com/business/companies/2021/06/how-social-media-companies-help-authoritarian-governments-censor-internet

Clifford, B. (2018) ““Trucks, Knives, Bombs, Whatever:” Exploring Pro-Islamic State Instructional Material on Telegram.” CTC Sentinel 11(5).

Cobler, N., & Herrera, S. (2018, August 13). “Bans don’t seem to be lessening reach of Alex Jones, InfoWars.” Statesman News Network, https://www.statesman.com/business/20180813/bans-dont-seem-to-be-lessening-reach-of-alex-jones-infowars.

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. Personality and Social Psychology Bulletin, 26(9), 1151–1164.

Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., & Ross, L. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. Journal of Personality and Social Psychology, 93(3), 415.

Coleman, P. (2021). The Way Out: How to Overcome Toxic Polarization. Columbia University Press.

Concha, J. (2020, July 1). “Trump dings CNN, 'Morning Joe' ratings as Tucker Carlson sets record,” The Hill: Media. https://thehill.com/homenews/media/505386-trump-dings-cnn-morning-joe-ratings-as-tucker-carlson-sets-record

Cook, J. (2017). Understanding and countering climate science denial. Journal and Proceedings of the Royal Society of New South Wales, 150.

Cook, J., Lewandowsky, S., & Ecker, U. K. (2017a). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. PloS One, 12(5), e0175799.

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017b). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence.(Research Article). PLoS ONE, 12(5), e0175799. https://doi.org/10.1371/journal.pone.0175799

Correll, J., Spencer, S. J., & Zanna, M. P. (2004). An affirmed self and an open mind: Self-affirmation and sensitivity to argument strength. Journal of Experimental Social Psychology, 40(3), 350–356.

Counter Extremism Project (2020, September 7). “Vehicles as Weapons of Terror.” https://www.counterextremism.com/vehicles-as-weapons-of-terror

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 2: Impediments to and strategies for change. BMJ Qual Saf, 22(Suppl 2), ii65–ii72.

Cruz, T. (2020, December 11). “Big Tech Believes There Is No Power That Can Constrain Them.” U.S. Senator for Texas, Ted Cruz. https://www.cruz.senate.gov/?p=press_release&id=5512

Culliford, E. (2019, December 12). “Facebook pledges $130 million to content oversight board, deakys naming members,” Reuters: Technology News. https://www.reuters.com/article/us-facebook-oversight/facebook-pledges-130-million-to-content-oversight-board-delays-naming-members-idUSKBN1YG1ZG

Currin, C., & Khaledi-Nasab, A. (2021). Depolarization of echo chambers by random dynamical nudge. ArXiv Preprint ArXiv:2101.04079.

Davis, P. K., Perry W. L., Brown, R. A., Yeung, D., Roshan, P. & Voorhies, P. (2013). Using Behavioral Indicators to Help Detect Potential Violent Acts: A Review of the Science Base. RAND Corporation.

Desmarais, S. L., Simons-Rudolph, J., Brugh, C. S., Schilling, E., & Hoggan, C. (2017). The state of scientific knowledge regarding factors associated with terrorism. Journal of Threat Assessment and Management, 4(4), 180–209. https://doi.org/10.1037/tam0000090

Díaz, A., & Hecht-Felella, L. (2021, August 4). Double Standards in Social Media Content Moderation. Brennan Centre for Social Justice, https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf

Dodgson, L. (2019). "YouTubers Have Identified a Long List of Words That Immediately Get Videos Demonetized, and They Include 'gay' and 'Lesbian' but Not 'Straight' or 'Heterosexual.'" Insider. https://www.insider.com/youtubers-identify-title-words-that-get-videos-demonetized-experiment-2019-10.

Dreyfuss, E. (2017). "Blaming the Internet For Terrorism Misses The Point." Wired. https://www.wired.com/2017/06/theresa-may-internet-terrorism/.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory & Cognition, 38(8), 1087–1100.

Eisenberger, N. I., Lieberman, M.D., & Williams, K. D. (2003). "Does Rejection Hurt? An FMRI Study of Social Exclusion." Science, 302(5643): 290–92. https://doi.org/10.1126/science.1089134.

Equal Access International. (2021). "Two Sides of the Same Coin? An Examination of Cognitive and Psychosocial Pathways Leading To Empowerment and Radicalization," EAI. https://www.equalaccess.org/resources/two-sides-of-the-same-coin-an-examination-of-the-cognitive-and-psychosocial-pathways-learning-to-empowerment-and-radicalization-and-a-model-for-reorienting-violent-radicalization/

European Commission. (2019, February 4). “Countering illegal hate speech online – EU Code of Conduct ensures swift response,” European Commission: Press Corner. https://ec.europa.eu/commission/presscorner/detail/en/IP_19_805

EUROPOL. (2019). “A Practical Guide to the First Rule of CTCVE.” Europol. https://www.europol.europa.eu/publications-documents/practical-guide-to-first-rule-of-ctcve

Facebook. (2020). “An Update to How We Address Movements and Organizations Tied to Violence.” About Facebook (blog). 2020. https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

------- (2020, August 11). “Our Commitment to Safety,” Facebook: Business. https://www.facebook.com/business/news/our-commitment-to-safety

------- (2021a). “Community Standards.” https://www.facebook.com/communitystandards/hate_speech.

------- (2021b). “Community Standards Enforcement | Transparency Center.” 2021. https://transparency.fb.com/data/community-standards-enforcement/

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? Personality and Social Psychology Bulletin, 41(12), 1665–1681.

Fino, A. (March, 2020). Defining Hate Speech: A Seemingly Elusive Task, Journal of International Criminal Justice, 18(1) March 2020, 31–57. https://doi.org/10.1093/jicj/mqaa023

Free Speech History (2019, February 23). “Episode 21 – The Bulwark of Liberty – Free Speech in 18th Century America, Part I .” Clear and Present Danger: A history of Free Speech, Free Speech History. http://www.freespeechhistory.com/2019/02/13/the-bulwark-of-liberty/

Frenkel, S. (2018, May 15). “Facebook Says It Deleted 865 Million Accounts, Mostly Spam,” The New York Times: Technology. https://www.nytimes.com/2018/05/15/technology/facebook-removal-posts-fake-accounts.html

Frenkel, S. (2020, December 18) “QAnon still spreading on Facebook, despite ban,” The New York Times: Technology. https://www.nytimes.com/2020/12/18/technology/qanon-is-still-spreading-on-facebook-despite-a-ban.html

Gebeily, M. (2021, May 10). “Instagram, Twitter blame glitches for deleting Palestinian posts.” Thomson Reuters Foundation: News, https://news.trust.org/item/20210510165535-38v1l/

Gill, P., Corner, E., Thornton, A., & Conway, M. (2015). What Are the Roles of the Internet in Terrorism? Measuring Online Behaviours of Convicted UK Terrorists.

Gilsinan, K. (2015). “ISIS and the ‘Internet Radicalization’ Trope.” The Atlantic. https://www.theatlantic.com/international/archive/2015/12/isis-internet-radicalization/419148/.

Gladstone, R., Goel, V. (2015, March 15) “ISIS is Adept on Twitter, Study Finds,” The New York Times: World https://www.nytimes.com/2015/03/06/world/middleeast/isis-is-skilled-on-twitter-using-thousands-of-accounts-study-says.html

Glenday, J. (2020, September 22). “Facebook, YouTube and Twitter advance hate speech talks with brands.” The Drum, https://www.thedrum.com/news/2020/09/23/facebook-youtube-and-twitter-advance-hate-speech-talks-with-brands

Goldberg, B. (2021, July 28). “Hate “Clusters” Spread Disinformation Across Social media. Mapping Their Networks Could Disrupt Their Reach.” Medium, https://medium.com/jigsaw/hate-clusters-spread-disinformation-across-social-media-995196515ca5

Google, Inc. (2021). Hate Speech Policy: YouTube Help. Accessed 1 October 2021. https://support.google.com/youtube/answer/2801939?hl=en

Granovetter, M. (1973). The Strength of Weak Ties, American Journal of Sociology 78(6), 1360-1380.

Hafez, M. & Mullins, C. (2015). The Radicalization Puzzle: A Theoretical Synthesis of Empirical Approaches to Homegrown Extremism. Studies in Conflict & Terrorism 38(11): 958–75. https://doi.org/10.1080/1057610X.2015.1051375.

Haidt, J. & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize. Social Justice Research 20(1): 98–116.

Haidt, J. (2013). The Righteous Mind: Why Good People Are Divided by Politics and Religion. New York: Vintage (Reprint edition).

Haque, U. (2019). "Big Tech's Big Frat-House Fascism Problem." Medium. https://eand.co/big-techs-big-frat-house-fascism-problem-fb4d010932a8.

Hamid, N. & Ariza, C. (2022). Offline Versus Online Radicalisation: Which is the Bigger Threat? GNET. https://gnet-research.org/wp-content/uploads/2022/02/GNET-Report-Offline-Versus-Online-Radicalisation.pdf

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. Communication Research, 39(6), 701–723.

Hassan, G., Brouillette-Alarie, S., Alava, S., Frau-Meigs, D., Lavoie, L., Fetiu, A., Varela, W., Borokhovski, E., Venkatesh, V., Rousseau, C., & Sieckelinck, S. (2018). Exposure to Extremist Online Content Could Lead to Violent Radicalization: A Systematic Review of Empirical Evidence. International Journal of Developmental Science, 12, 71–88. https://doi.org/10.3233/DEV-170233

Hennes, E. P., Nam, H. H., Stern, C., & Jost, J. T. (2012). Not all ideologies are created equal: Epistemic, existential, and relational needs predict system-justifying attitudes. Social Cognition, 30(6), 669–688.

Hicks, P. (2021, July 14). Online content moderation and internet shutdowns [Speech Transcript]. United Nations Human Rights Office, https://www.ohchr.org/Documents/Press/Press%20briefing_140721.pdf

Horgan, J. (2008). From Profiles to Pathways and Roots to Routes: Perspectives from Psychology on Radicalization into Terrorism. The ANNALS of the American Academy of Political and Social Science 618 (1): 80–94.

------ (2012). Discussion Point: The End of Radicalization. National Consortium for the Study of Terrorism and Responses to Terrorism, 28.

Horwitz, J. (2021, September 13). "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt." Wall Street Journal: The Facebook Files,

https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353?mod=article_inline

Horwitz, J. & Seetharaman, D. (2020, May 26). "Facebook Executives Shut Down Efforts to Make the Site Less Divisive." Wall Street Journal:Tech. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499.

Isaac, M. (2016, February 6). "Twitter Steps Up Efforts to Thwart Terrorist Tweets," The New York Times: Technology https://www.nytimes.com/2016/02/06/technology/twitter-account-suspensions-terrorism.html

Jay, S. (2007). The Creation of the First Amendment Right to Free Expression: From the Eighteenth Century to the Mid-Twentieth Century. WM. Mitchell L. Rev., 34, 773.

Jenkins, M. M., & Youngstrom, E. A. (2016). A randomized controlled trial of cognitive debiasing improves assessment and treatment selection for pediatric bipolar disorder. Journal of Consulting and Clinical Psychology, 84(4), 323.

Johnson, N.F., Zheng, M., Vorobyeva, Y., Gabriel, A., Qi, H., Velásquez, N., Manrique, P., Johnson, D., Restrepo, E. & Song, C. (2016). New Online Ecology of Adversarial Aggregates: ISIS and Beyond. Science, 352(6292): 1459–63.

Johnson, Nicola F., R. Leahy, N. Johnson Restrepo, Nicolas Velasquez, Ming Zheng, P. Manrique, P. Devkota, and Stefan Wuchty. 2019. "Hidden Resilience and Adaptive Dynamics of the Global Online Hate Ecology." Nature 573 (7773): 261–65.

Jones, SG., Vallee, C., Newlee, D., Harrington, N., Sharb, C., & Byrne, H. (2018). The Evolution of the Salafi-Jihadist Threat. Centre for Strategic and International Studies https://www.csis.org/analysis/evolution-salafi-jihadist-threat.

Jong, J., Whitehouse, H., Kavanagh, C. & Lane, J. (2015). Shared Negative Experiences Lead to Identity Fusion via Personal Reflection. PLoS ONE 10(12).

Joyella, M. (2021, July 27). "Fox News Dominates July Cable News Ratings As All Networks See Declines," Forbes: Media.

https://www.forbes.com/sites/markjoyella/2021/07/27/fox-news-dominates-july-cable-news-ratings-as-all-networks-see-declines/

Jozwiak, M. (2018, March 15). Internet, Freedom of Speech and Slippery Slope Argument – The Case of the 'Right to Be Forgotten' http://dx.doi.org/10.2139/ssrn.3141370

Kaiser, J. (2021, June 15). Deplatforming the far-right: An analysis of YouTube and BitChute. Berkman Klein Centre for Internet & Society, https://cyber.harvard.edu/story/2021-06/deplatforming-far-right-analysis-youtube-and-bitchute

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. Scientific reports, 6(1), 1-11.

Kari, P. (2021, February 4). "'It Let White Supremacists Organize': The Toxic Legacy of Facebook's Groups." The Guardian: Technology. http://www.theguardian.com/technology/2021/feb/04/facebook-groups-misinformation

Kaufman, S. B. (2020, June 29). "Unraveling the Mindset of Victimhood." Scientific American. https://www.scientificamerican.com/article/unraveling-the-mindset-of-victimhood/

Kenyon, J., Binder, J., & Baker-Beall, C. (2021) Exploring the role of the Internet in radicalisation and offending of convicted extremists. HM Prison & Probation Service: Ministry of Justice Analytical Series, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1017413/exploring-role-internet-radicalisation.pdf

Khaffary, Shirin. (2022) "Does banning extremists online work? It depends." Vox https://www.vox.com/recode/22913046/deplatforming-extremists-ban-qanon-proud-boys-boogaloo-oathkeepers-three-percenters-trump

Kimball, S. (2019, March 30). Zuckerberg backs stronger Internet privacy and election laws: "We need a more active role for governments." CNBC. https://www.cnbc.com/2019/03/30/mark-zuckerberg-calls-for-tighter-internet-regulations-we-need-a-more-active-role-for-governments.html

King Jr, M.L. (2019). Strength to Love. Beacon Press.

Kow, C.S., Merchant, H.A., Mustafa, Z.U., & Hasan, S.S. (2021). The association between the use of ivermectin and mortality in patients with COVID-19: a meta-analysis. Pharmacological Reports, 73, 1473-1479. DOI: 10.1007/s43440-021-00245-z

Kteily, N., Hodson, G. & Bruneau, E. (2016). They See Us as Less than Human: Metadehumanization Predicts Intergroup Conflict via Reciprocal Dehumanization. Journal of Personality and Social Psychology 110(3): 343.

Lane, J. E. (2021). Understanding Religion Through Artificial Intelligence: Bonding and Belief. Bloomsbury Publishing.

Lane, J., Shults, F., & Wildman, W. (2018). A potential explanation for self-radicalisation. Behavioral and Brain Sciences, 41, E207. doi:10.1017/S0140525X18001760

Lane, J. E., McCaffree, K. & LeRon Shults, F. (2021). Is Radicalization Reinforced by Social Media Censorship? ArXiv Preprint ArXiv:2103.12842.

Langvardt, K. (2017, AUgust 1) Regulating Online Content Moderation. Georgetown Law Journal, 106(5) http://dx.doi.org/10.2139/ssrn.3024739

Lawrence, J.M., Meyerowitz-Katz, G., Heathers, J.A.J., Brown, N.J.L. & Sheldrick, K.A. (2021). The lesson of ivermectin: meta-analyses based on summary data alone are inherently unreliable. Nature Medicine. DOI: 10.1038/s41591-021-01535-y

Lee, D. (2018, May 15). "Facebook details scale of abuse on its site," BBC: Technology. https://www.bbc.com/news/technology-44122967

Lee, B., & Knott, K. (2020). More Grist to the Mill? Reciprocal Radicalisation and Reactions to Terrorism in the Far-Right Digital Milieu. Perspectives on Terrorism, 14(3), 98-115.

Leibniz, G. W. (2009). The Philosophical Works of Leibnitz. Рипол Классик.

Lennard, N. (2020). "Facebook's Ban on Far-Left Pages Is an Extension of Trump Propaganda." The Intercept (blog). https://theintercept.com/2020/08/20/facebook-bans-antifascist-pages/.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction. Psychological Science in the Public Interest, 13(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. Journal of Applied Research in Memory and Cognition, 6(4), 353–369.

Locke, John. 1690. “The Project Gutenberg EBook of Second Treatise Of Government By John Locke.” https://www.gutenberg.org/files/7370/7370-h/7370-h.htm.

Luzsa, R., & Mayr, S. (2021). False consensus in the echo chamber: Exposure to favorably biased social media news feeds leads to increased perception of public support for own opinions. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 15(1).

Magid, L. (2018, May 15). “Facebook Reports Numbers on Nudity, Sex, Violence, Hate Speech And Other Banned Content.” Forbes, https://www.forbes.com/sites/larrymagid/2018/05/15/facebook-reports-numbers-on-nudity-sex-violence-hate-speech-and-other-banned-content/?sh=669cb36b2968

McCauley, C. (2020). The Essence of Hate and Love. In R. J. Sternberg (Ed.) Perspectives on Hate: How It Originates, Develops, Manifests, and Spreads. Washington, D.C.: APA Books.

McCauley, C., & Moskalenko, S. (2010). Individual and Group Mechanisms of Radicalization. In Fenstermarcher, Kuznar & Speckhard (Eds.). Protecting the Homeland from International and Domestic Terrorism Threats: Current Multi-Disciplinary Perspectives on Root Causes, the Role of Ideology and Programs for Counter-Radicalization and Disengagement. Multi-Disciplinary White Papers in Support of Counter-Terrorism and Counter-WMD: Multi-Agency and Airforce Research Laboratory.

McCauley, C., & Moskalenko, S. (2017). Understanding political radicalization: The two-pyramids model. American Psychologist, 72(3), 205.

McEwan, S. (2017). Nation of Shitposters: Ironic Engagement with the Facebook Posts of Shannon Noll as Reconfiguration of an Australian National Identity. PLATFORM: Journal of Media & Communication 8(2).

McGregor, H. A., Lieberman, J. D., Greenberg, J., Solomon, S., Arndt, J., Simon, L., & Pyszczynski, T. (1998). Terror Management and Aggression: Evidence That Mortality Salience Motivates Aggression Against Worldview-Threatening Others. Journal of Personality and Social Psychology, 74(3), 590–605. https://doi.org/10.1037/0022-3514.74.3.590

Mitchell, A., Gottfried, J., Kiley, J., Matsa, KE.(2014, October 21). “Political Polarization & Media Habits - Social Media: Conservatives More Likely to Have Like-Minded Friends.” Pew Research Centre, https://www.pewresearch.org/journalism/2014/10/21/political-polarization-media-habits/#social-media-conservatives-more-likely-to-have-like-minded-friends

Mitchell, A., & Walker, M. (2021, August 18). “More Americans now say government should take steps to restrict false information online than in 2018.” Pew Research Centre. https://www.pewresearch.org/fact-tank/2021/08/18/more-americans-now-say-government-should-take-steps-to-restrict-false-information-online-than-in-2018/

Molenberghs, P., & Louis, WR. (2018) Insights From fMRI Studies Into Ingroup Bias. Frontiers in Psychology. 9(1868). doi: 10.3389/fpsyg.2018.01868

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions. Policy Insights from the Behavioral and Brain Sciences, 2(1), 129–140. https://doi.org/10.1177/2372732215600886

Morrison, S. “Facebook and Twitter made special world leader rules for Trump. What happens now?” Vox: Recode. https://www.vox.com/recode/22233450/trump-twitter-facebook-ban-world-leader-rules-exception

Moskalenko, S., & McCauley, C. (2020). Radicalization to Terrorism: What Everyone Needs to Know. Oxford University Press.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. (2021). Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–13.

Mueller, J., & Stewart, M.G. (2021). Terrorism and Bathtubs: Comparing and Assessing the Risks. Terrorism and Political Violence 33(1): 138–63. https://doi.org/10.1080/09546553.2018.1530662.

National Institute for Justice. (2015). Radicalization and Violent Extremism: Lessons Learned from Canada, the U.K. and the U.S. U.S. Department of Justice. https://www.ncjrs.gov/pdffiles1/nij/249947.pdf.

Nelson, S. (2021, July 15). “White House ‘Flagging’ Posts for Facebook to Censor over COVID ‘Misinformation.’” New York Post (blog). https://nypost.com/2021/07/15/white-house-flagging-posts-for-facebook-to-censor-due-to-covid-19-misinformation/.

Neumann, P. R. (2003). “The Trouble with Radicalization.” International Affairs 89(4): 873–93.

Newton, C.(2019, February 25) “The Trauma Floor: The Secret Lives of Facebook Moderators in America,” The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Nicas, J. (2018, September 4). “Alex Jones Said Bans Would Strengthen Him. He Was Wrong,” The New York Times: Technology. https://www.nytimes.com/2018/09/04/technology/alex-jones-infowars-bans-traffic.html

Nicas, J., Isaac, M., Frenkel, S. (2021, January 14) “Millions Flock to Telegram and Signal as Fears Grow Over Big Tech.” The New York Times: Technology. https://www.nytimes.com/2021/01/13/technology/telegram-signal-apps-big-tech.html

O’Neill, KF. (2009). “True Threats,” The First Amendment Encyclopedia. (Updated June 2017 by David L. Hudson Jr.) https://www.mtsu.edu/first-amendment/article/1025/true-threats

Packingham v. North Carolina: Syllabus. No. 15–1194 (U.S. Supreme Court, 2016). https://www.supremecourt.gov/opinions/16pdf/15-1194_08l1.pdf

Padhy, B.M., Mohanty, R.R., Das, S., Meher, B.R. (2020). Therapeutic potential of ivermectin as add-on treatment in COVID 19: A systematic review and meta-analysis. Journal of Pharmacy & Pharmaceutical Sciences. 23: 357-495.

Paul, K. (2019, August 9). 8chan: Ex-users of far-right site flock to new homes across internet. The Guardian. https://www.theguardian.com/us-news/2019/aug/08/8chan-shutdown-users-social-media

Paul, Kari. 2021. “‘It Let White Supremacists Organize’: The Toxic Legacy of Facebook’s Groups.” The Guardian. 2021. https://www.theguardian.com/technology/2021/feb/04/facebook-groups-misinformation

Pauwels, L. JR., & Heylen, B. (2020). Perceived Group Threat, Perceived Injustice, and Self-Reported Right-Wing Violence: An Integrative Approach to the Explanation of Right-Wing Violence. Journal of Interpersonal Violence 35(21–22): 4276–4302.

Popper, K. R. (1966). The Open Society and Its Enemies.(Rev. 2 Vols.). Princeton: Princeton University Press.

Pretus, C., Hamid, N., Sheikh, H., Ginges, J., Tobeña, A., Davis, R., Vilarroya, O., & Atran, S. (2018). Neural and Behavioral Correlates of Sacred Values and Vulnerability to Violent Extremism. Frontiers in psychology, 9, 2462. https://doi.org/10.3389/fpsyg.2018.02462

Rampton, R., & Shepardson, D. (2019, July 11). "Trump rips tech firms at 'free speech' summit," Reuters: Media & Telecoms. https://www.reuters.com/article/us-usa-socialmedia/trump-rips-tech-firms-at-free-speech-summit-idUSKCN1U6189

Rathje, S., Van Bavel, J.J., & van der Linden, S. (2021). Out-Group Animosity Drives Engagement on Social Media. Proceedings of the National Academy of Sciences 118(26). https://doi.org/10.1073/pnas.2024292118.

Rauchfleisch, A., & Kaiser, J. (2021). Deplatforming the Far-Right: An Analysis of YouTube and BitChute. Rochester, NY: Social Science Research Network (SSRN Scholarly Paper ID 3867818). https://doi.org/10.2139/ssrn.3867818.

Reagan, B., Horwitz, J., Scheck, J., Seetharaman, D., Wells, G. (2021, September 20). "The Facebook Files." Wall Street Journal, https://www.wsj.com/articles/the-facebook-files-11631713039?redirect=amp#click=https://t.co/P0jJHS9e9H

Reed, A., & Ingram, H. (2018). A Practical Guide to the First Rule of CTCVE. Europol. 2018. https://www.europol.europa.eu/publications-documents/practical-guide-to-first-rule-of-ctcve

Reidy, K. (2018) Radicalization as a Vector: Exploring Non-Violent and Benevolent Processes of Radicalization, Journal for Deradicalization, 14, 1-46.

Reidy, K. (2019). Benevolent Radicalization. Perspectives on Terrorism 13(4): 1–13.

Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & West, R. (2020). Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. arXiv:2010.10397

Roberts, ST. (2018, June 25). "Meet the people who scar themselves to clean up our social media networks," MacLean's: Opinion. https://www.macleans.ca/opinion/meet-the-people-who-scar-themselves-to-clean-up-our-social-media-networks/

Rosen, G. (2021). "Our Response to the Violence in Washington - About Facebook." https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/.

Rui F., Xu, K., & Zhao, J. (2018). "Higher Contagion and Weaker Ties Means Anger Spreads Faster than Joy in Social Media." ArXiv Preprint ArXiv:1608.03656.

Sageman, M. (2015). "Latif, et al. v. Holder, et al. - Declaration of Marc Sageman." American Civil Liberties Union. https://www.aclu.org/legal-document/latif-et-al-v-holder-et-al-declaration-marc-sageman.

Saltman, E. (2021, July 11). "Challenges in Combating and Extremism Online," Lawfare Blog. https://www.lawfareblog.com/challenges-combating-terrorism-and-extremism-online

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2021). Social influence and unfollowing accelerate the emergence of echo chambers. Journal of Computational Social Science, 4(1), 381–402.

Scheck, J., Purnell, N., & Horwitz, J.(2021, September 16). "Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show." Wall Street Journal: The Facebook Files, https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953?mod=article_inline

Schuurman, B., & Taylor, M. (2018). Reconsidering Radicalization: Fanaticism and the Link Between Ideas and Violence, Perspectives on Terrorism 12(1), 1-22.

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? In Psychology of learning and motivation, 41, 265–292.

Shehabat, A., Mitew, T., & Alzoubi, Y. (2017). "Encrypted Jihad: Investigating the Role of Telegram App in Lone Wolf Attacks in the West." Journal of Strategic Security 10(3), 27-53. DOI: http://doi.org/10.5038/1944-0472.10.3.1604

Sheikh, H., Ginges, J., Coman, A., Atran, S. (2012) Religion, group threat and sacred values. Judgement and Decision Making 7(2), 110-118.

Shepardson, D. (2019, April 10) "Facebook, Google accused of anti-conservative bias at U.S. Senate hearing." Reuters. https://www.reuters.com/article/us-usa-congress-socialmedia-idUSKCN1RM2SJ

Shults, F. L. (2018). "Can We Predict and Prevent Religious Radicalization?" In Øverland, G. (Ed.) Violent Extremism in the 21st Century: International Perspectives. (pp 45–71). Cambridge: Cambridge Scholars Press.

Shults, F. L. (2020). Toxic theisms? New strategies for prebunking religious belief-behavior complexes. Journal of Cognitive Historiography, 5(2), 1–19. https://doi.org/DOI: 10.1558/jch.38074

Shults, F. L., & Gore, R. (2020). Modeling Radicalization and Violent Extremism. Advances in Social Simulation, 405–10. Springer, Cham.

Shults, F. L., Gore, R., Wildman, W. J., Lynch, C., Lane, J. E., & Toft, M. (2018). A Generative Model of the Mutual Escalation of Anxiety Between Religious Groups. Journal of Artificial Societies and Social Simulation, 21(4), DOI: 10.18564/jasss.3840.

Shults, F. L., Lane, J. E., Diallo, S., Lynch, C., Wildman, W. J., & Gore, R. (2018). Modeling Terror Management Theory: Computer Simulations of the Impact of Mortality Salience on Religiosity. Religion, Brain & Behavior, 8(1), 77–100.

Shuster, S. (2021, January 18). "'Everyone Thinks I'm a Terrorist': Capitol Riot Fuels Calls for Domestic War on Terror." TIME Magazine. https://time.com/5930592/everyone-thinks-im-a-terrorist-capitol-riot-fuels-calls-for-domestic-war-on-terror/

Similar Web. (August 2021). "Traffic Overview: Infowars.com" SimilarWeb, https://www.similarweb.com/website/infowars.com/#overview

Singh, S. (2019, July 22) Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content. New America Foundation: Open Technology Institute. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

Sjøen, M. M., & H. Jore, S.H. (2019). Preventing Extremism through Education: Exploring Impacts and Implications of Counter-Radicalisation Efforts. Journal of Beliefs & Values 40(3), 269–83.

Smith, A., & Anderson, M. (2018, March 1). "Social Media Use 2018: Demographics and Statistics." Pew Research Center: Internet, Science & Tech (blog). https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/.

Song, Y., Dai, X.Y., & Wang, J. (2016). Not All Emotions Are Created Equal: Expressive Behavior of the Networked Public on China's Social Media Site. Computers in Human Behavior, 60: 525–33.

Spencer, N. (2021, August 10). "Alt-Tech Social Networks: What Investigators and Analysts Need to Know," Loss Prevention Magazine. https://losspreventionmedia.com/alt-tech-social-networks/

Statt, N. (2020, May 26). "Facebook reportedly ignored its own research showing algorithms divided users," The Verge. https://www.theverge.com/2020/5/26/21270659/facebook-division-news-feed-algorithms

Stephens, G.M. (n.d.). "John Locke: His American and Carolinian Legacy." John Locke Foundation. Accessed July 17, 2021. https://www.johnlocke.org/john-locke-his-american-and-carolinian-legacy/.

Swann, W.B., Jetten, J., Gómez, A., Whitehouse, H., & Bastian, B. (2012). When Group Membership Gets Personal: A Theory of Identity Fusion. Psychological Review, 119(3): 441–56. https://doi.org/10.1037/a0028589.

Streisand v. Adelman, SC 077 257. (CA Sup. Ct., 2003) https://www.californiacoastline.org/streisand/slapp-ruling-tentative.pdf

Taylor, J. (2020, June 17). "Not Just Nipples: How Facebook's AI Struggles to Detect Misinformation." The Guardian: Technology.

http://www.theguardian.com/technology/2020/jun/17/not-just-nipples-how-facebooks-ai-struggles-to-detect-misinformation.

Taylor, M. (2010). Is Terrorism a Group Phenomenon? Aggression and Violent Behavior 15(2), 121–29.

Taylor, M., & Currie, P.M. (2012). Terrorism and Affordance. A&C Black.

Tech Transparency Project. (2020, August 12). Fecbook's Boogaloo Problem: A record of failures. Tech Transparency Project.

Tech Transparency Project. (2021). Capitol Attack Was Months in the Making on Facebook. Tech Transparency Project. https://www.techtransparencyproject.org/articles/capitol-attack-was-months-making-facebook

Terdiman, Daniel. (2018). Here's How Facebook Uses AI to Detect Many Kinds of Bad Content. FastCompany. Published 2 May 2018. https://www.fastcompany.com/40566786/heres-how-facebook-uses-ai-to-detect-many-kinds-of-bad-content

Toft, M.D., Philpott, D., & Shah, T. S. (2011). God's Century: Resurgent Religion and Global Politics. Norton.

Trenchard, J. (1723, January 5) "Of Liberty and Necessity," Letter No. 110. In Trenchard, J. & Gordon, T. "Cato's Letters, or Essays on Liberty, Civil and Religious, and Other Important Subjects (Selections)," Natural Law, Natural Rights and American Constitutionalism: Early Modern Liberal Roots of Natural Law. https://www.nlnrac.org/earlymodern/radical-whigs-and-natural-rights/documents/cato-letters#cato_libnec

Troianovski, A., & Nechepurenko, I. (2021, September 19) "Russian Election Shows Declining Support for Putin's Party." The New York Times: Europe, https://www.nytimes.com/2021/09/19/world/europe/russia-election-google.html

Trump, D.J. (2021, July 8) "Why I'm Suing Big Tech." Wall Street Journal. https://www.wsj.com/articles/donald-j-trump-why-im-suing-big-tech-11625761897

Twitter (2016, February 5). “Combating Violent Extremism.” Twitter (blog). https://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html

Twitter. (2021). “Twitter’s Policy on Hateful Conduct | Twitter Help.” https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

Twitter Inc. (2021, January 8). “Permanent suspension of @realDonaldTrump.” Twitter: Blog, https://blog.twitter.com/en_us/topics/company/2020/suspension

Tworek, H. (2021, January 13). “The Dangerous Inconsistencies of Digital Platform Policies.” Centre for International Governance Innovation. https://www.cigionline.org/articles/dangerous-inconsistencies-digital-platform-policies/

UNESCO. (2017). “Youth and Violent Extremism on Social Media: Mapping the Research - UNESCO Digital Library.” https://unesdoc.unesco.org/ark:/48223/pf0000260382.

Varvin, Sverre. (2005). Humiliation and the Victim Identity in Conditions of Political and Violent Conflict. The Scandinavian Psychoanalytic Review 28(1): 40–49.

Velásquez, N., Leahy, R., Restrepo, N.J., Lupu, Y.; Sear, R., Gabriel, R., Jha, OK., Goldber, B., Johnson, NF. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. Scientific Reports 11, 11549. https://doi.org/10.1038/s41598-021-89467-y

Vengattil, M., & Dave, P. (2019, February 28). “Some Facebook content reviewers in India complain of low pay, high pressure,” Reuters. https://www.reuters.com/article/us-facebook-content-india-feature/some-facebook-content-reviewers-in-india-complain-of-low-pay-high-pressure-idUSKCN1QH15I

Vidino, L. (2010) Countering Radicalization in America: Lessons from Europe, United States Institute for Peace (USIP), Special Report.

Villasenor, J. (2017). “Views among College Students Regarding the First Amendment: Results from a New Survey.” Brookings (blog). https://www.brookings.edu/blog/fixgov/2017/09/18/views-among-college-students-regarding-the-first-amendment-results-from-a-new-survey/.

Violent Islamist extremism, 2007 [electronic resource]: hearings before the Committee on Homeland Security and Governmental Affairs, United States Senate, One Hundred Tenth Congress, first session. (2009). U.S. G.P.O. (Testimony of Brian Jenkins)

Volz, D. (2020, October 8). "Facebook Takes Down Network Tied to Conservative Group, Citing Fake Accounts." Wall Street Journal: Tech. https://www.wsj.com/articles/facebook-takes-down-network-tied-to-conservative-group-citing-fake-accounts-11602174088.

Von Behr, I. (2013). "Radicalisation in the Digital Era: The Use of the Internet in 15 Cases of Terrorism and Extremism." https://apo.org.au/node/36281

Ward, K. D. (1997). Free Speech and the Development of Liberal Virtues: An Examination of the Controversies Involving Flag-Burning and Hate Speech. U. Miami L. Rev., 52(733).

Wells, G., Horwitz, J., Seetharaman, D. "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show." Wall Street Journal: The Facebook Files, https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?mod=hp_lead_pos7&mod=article_inline

Whitbourne, S. K. (2010). "In-Groups, out-Groups, and the Psychology of Crowds." Psychology Today, 181–203.

Wike, R., & Simmons, K. (2015). "Global Support for Principle of Free Expression, but Opposition to Some Forms of Speech." Pew Research Center, 18.

Wille, B. (2020, September 10). "Video Unavailable": Social Media Platforms Remove Evidence of War Crimes. Human Rights Watch. https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes

Wiliams, P. (2021, March 13). "Still a Mystery: Was the Capitol Riot Planned Far in Advance?" NBC News. https://www.nbcnews.com/news/us-news/still-mystery-was-capitol-riot-planned-far-advance-n1261020

Wiltz, T. (2017, January 13). "Should Social Media Be Banned in Prison?" http://pew.org/2ilo8KG.

Wong, Q., & Morse, A. (2021, February 16). “Parler returns online after monthlong absence: Here's what you need to know,” CNet: News. https://www.cnet.com/news/parler-returns-online-after-month-long-absence-heres-what-you-need-to-know/

Yoo, J. (2017). Ideological Homophily and Echo Chamber Effect in Internet and Social Media. Student International Journal of Research, 4(1). http://www.sijr.ac/wp-content/uploads/2017/02/Ideological-Homophily-and-Echo-Chamber-Effect.pdf

York, J.C., & McSherry, C. (2019, April 29). “Content Moderation is Broken. Let Us Count the Ways.” Electronic Frontier Foundation. https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways

YouTube. (2021). “Hate Speech Policy - YouTube Help.” https://support.google.com/youtube/answer/2801939.

Yurieff, K., Fung, B., & O'Sullivan, D. (2021, January 10). “Parler: Everything you need to know about the banned conservative social media platform.” CNN: Business Tech, https://www.cnn.com/2021/01/10/tech/what-is-parler/index.html