For(e)Dialogue • Vol 4: Special Issue 1: Media and the Far-Right

'Censorship-free' platforms: Evaluating content moderation policies and practices of alternative social media

Nicole Buckley[1], **Joseph S. Schafer**[1]

[1]**University of Washington**

Published on: Feb 03, 2022

Nicole Buckley, University of Washington, nicolehb@uw.edu

Joseph S. Schafer, University of Washington, schaferj@uw.edu

Abstract

Following the development and implementation of mainstream social media platforms' election-related speech policies, a renewed wave of criticism emerged from the U.S. ideological right. Several months before the 2020 U.S. presidential election, conservative politicians, pundits, and self-described patriots alleged that their speech was being censored by "Big Tech." This resulted in right-leaning influencers, and many of their followers, migrating to alternative online platforms to avoid moderation. Alternative social media, such as Parler, Bitchute, Gab, and Gettr, describe themselves as unmoderated hubs for "free speech," signalling an invitation for users to voice everything from unpopular opinions, to misinformation, to hate speech. Yet when pushed by technology infrastructure platforms like Apple's App Store and Google's Play Store to address missing or substandard moderation practices, "alt-tech" platforms were forced to create or adapt ad hoc, often minimalistic, content moderation policies.

Our research explores and evaluates these policies in comparison to mainstream platforms, and analyses how moderation policies interact with the ideological framework asserted at an alternative platform's nascence. Our work provides necessary insight into the potential motivations for one potential source of U.S. internet platform oversight. With few immediately available regulatory options, assessing the viability of alternatives is crucial. This is particularly true as severe legislative gridlock stalls meaningful reform to the federal law perhaps most capable of improving platforms' moderation practices. Because private regulation appears to be the most immediate solution to address new breeding grounds for mis- and disinformation, inquiry into alternative platforms' adoption and enforcement of moderation policies is needed. Our paper concludes with questions for future research into the efficacy of alternative platforms' policy implementation; it is imperative to distinguish legitimate moderation from mere shells constructed to retain profit in parallel with ideological posturing.

Keywords: Content Moderation, Social Media, Platform Policy, Section 230, Misinformation

1. Introduction

Online conversation casts a broad net. Social media has become integral to understanding an array of social issues and events. In fact, recent research shows online behaviour is relevant to conversations around election misinformation (Center for an Informed Public et al., 2021), social activism (Arif et al., 2018; Jackson et al., 2018; Vis et al., 2019), crisis events (Starbird et al., 2014; Vieweg et al., 2010), and public health (Koltai, 2020; Papakyriakopoulos et al., 2020). Calls to understand the totality of the digital ecosystem grow, and this paper works toward that end. We focus on "alt-tech" platforms (smaller, conservative-leaning platforms with less stringent moderation practices) rather than dedicating an exclusive focus to well-researched and heavily scrutinised popular social media (Freelon et al., 2020; Gillespie et al., 2020).

Within the context of U.S. elections, studying "alt-tech" platforms is particularly important. As we found at the Election Integrity Partnership, these spaces often host and rarely remove blatant hate speech and misinformation (Center for an Informed Public et al., 2021).[1] Detailing "alt-tech" platforms' steadfast dedication to hosting objectionable content underscores the growing possibility for real-world impact. For example, "alt-tech" platform Parler was considered a major player in the January 6 attack on the U.S. Capitol for its failure to moderate what became successful extremist organising efforts (Heilweil & Ghaffary, 2021; Shepardson, 2021). Although Parler was sanctioned by some actors for the role it played in facilitating the attack, it ultimately returned to major digital marketplaces with minimally changed policies. Deepening understanding about how and why "alt-tech" platforms maintain a calculated hesitation to moderate, therefore, is a primary component in the machinery needed to prevent similar violent, real-world events in the future.

Platform policies are an essential component of how platforms define themselves and shape norms around their purpose and use (Gillespie, 2010). Therefore, one way that we can gain an improved understanding of the divide between "Big Tech" and "alt-tech" platforms is through comparing platforms' policies.[2] While far from the only artefact that defines a platform (Pfaffenberger, 1992), platform policies are nevertheless instrumental in shaping the interactions that a platform has both with its users and with other entities (such as digital marketplaces, where application developers and software users can interact with, purchase, or create products online).

Some level of moderation is always necessary to comply with a given set of laws. In France, new legislation protects against specific types of "information manipulation" and provides legal injunctions for particularly egregious circumstances (Le

Gouvernement francais, 2018). In Germany, pro-Nazi and neo-Nazi content is explicitly banned online and off (Stegbauer, 2007). In the United States, Child Sexual Abuse Material (CSAM) is illegal to produce and distribute on the Internet (National Center for Missing and Exploited Children, n.d.). Nonetheless, regulating speech online is a difficult task; because platforms are internationally accessible (with users, offices, and hardware spread across the globe), they are often beyond the reach of any one nation's legal framework. Therefore, gaining a unified understanding about the necessary boundaries and protections surrounding online speech is practically impossible. However, platforms' own policies project one set of guidelines that define the universe of available topics about which platform users can post. Conversely, platform policies have also allowed certain types of objectionable, perhaps unprotected, speech to thrive. In the 1990s, public outcry in the United States swirled around the availability of pornographic or extremely violent content online (Citron & Wittes, 2017). More recently, widespread alarm about hate speech and misinformation surfaced around the 2020 U.S. election (Bazelon, 2020; Hill & Freelon, 2020). Still, where one platform might strictly regulate hate speech, another may be silent. Understanding how platform policies evolve and regulate speech is critical to earnestly hypothesising about the opinions of the actors behind those platforms—and what *they* consider important and acceptable. Put differently, the degree to which any platform policy regulates speech is a signal to users and marketplaces about an aspect of the "platform's" own ideology.

Because platform policies are relevant to understanding both the "Big Tech" / "alt-tech" divide and how policies help inform inter- and intra-platform interactions, we sought to systematically study policy differences between platform types. In particular, we focused on (1) if there were notable differences in speech policy language across "alt-tech" and "Big Tech" platforms; and (2) whether digital marketplaces reacted to available policy or actual content.

Our research shows that, despite many similarities, there are several notable policy differences between "Big Tech" and "alt-tech" platforms. This paper explores four dominant policy differences relating to (1) hate speech, (2) sceptical-veracity content, (3) policy exceptions, and (4) policy clarity and transparency. We argue that digital marketplaces do not treat "Big Tech" and "alt-tech" platforms similarly despite a similar prevalence of sanctionable content across both types of platforms, which may be partially attributable to a significant focus on policy language over platform content, among other considerations.

We begin with a brief background discussion that focuses on platform behaviour and user ideology. After describing the landscape in which platforms create and implement policy, we introduce our methods, which consist of a qualitative content analysis of two pairs of mainstream and alternative platforms (Hsieh & Shannon, 2005; Mayring, 2014). Next, we display the content analysis results and present a discussion about its most salient attributes. Finally, we conclude by discussing how the similarities and differences between mainstream and alternative platforms interact with a variety of digital marketplaces, and present areas for future research.

2. Background

2.1 Defining Platforms

Being described as a platform is itself a non-neutral definition, as this helps to frame what a service provides, its aims, and its values (Gillespie, 2010). Traditionally, the label "platform" conjures images of openness and ideological neutrality (Gillespie, 2010). Social media corporations use a variety of techniques in order to associate particular connotations with their business; these include the affordances platforms provide their users and the documents and official statements platforms make. Platform policies about user content are therefore important for two reasons. First, the policies a platform publishes represent a crucial set of public statements which help define what content the platform views as acceptable for their digital ecosystem. Second, the content policies can be thought of as a description of affordances, since content which does not violate a platform's policies is content that users are afforded the ability to discuss.

Certainly, platform policy documents are not the only components capable of defining platforms. A platform's technical affordances, the content and behaviour of its users, the norms which its communities develop, and more are also aspects of how a particular platform is determined to be a platform and distinguished from its competitors. Such forces often act concurrently, and frequently push against each other, as competing actors each seek to define the platform in ways beneficial to themselves (Pfaffenberger, 1992). However, prior research demonstrates that content moderation policies have significant impacts on the tone and topics of discussions in online spaces; therefore, we chose to focus on this particular force—policy—shaping social media platforms (Gibson, 2019).

Because policies are an essential part of this definitional process, we train our attention on those belonging to Twitter, YouTube, Parler, and Bitchute. The first two

embody what platforms are generally: gathering spaces for the public to virtually engage on a variety of topics and through a variety of mediums. The latter two are alternatives to the mainstream; they are developed with certain users in mind and host a relatively smaller amount of content trained on a specific attitude. We compare the platforms via a "Big Tech" versus "alt-tech" analogue, first coupling together "Big Tech" YouTube and "alt-tech" Bitchute. The two platforms are placed against one another primarily due to similarity in composition and affordances. Bitchute and YouTube both contain the same general aesthetic in terms of layout. Both Bitchute and YouTube centre longer-form, video-based content (presented horizontally and without a vertical "scrolling" feature). Additionally, both maintain similar formats for user interactions with user-posted content: comments located underneath videos. Likewise, we compared Twitter and Parler based on their overall similarity. Parler is a microblogging platform similar to Twitter, designed as an alternative for those ousted from or unimpressed with Twitter. Both platforms again contain the same general layout and prioritise short, text-based posts.

2.2 Evolving Standards of Platform Behaviour

Over the last several years, social media platforms have faced increasing public pressure to moderate content, as issues of problematic content like online harassment, hate speech, and misinformation have become increasingly salient in public discourse (Anti-Defamation League et al., 2020; De Vynck & Lerman, 2021; Frenkel, 2021). Platforms responded in multiple ways in attempts to improve their content moderation practices. This included new oversight mechanisms (Klonick, 2020), new policies on acceptable forms of user-generated content, and changes to consequences for creating standards-violating content (Center for an Informed Public et al., 2021). This contrasts with prior platform goals which made moderation practices as invisible and unobtrusive as possible (Roberts, 2016).

In addition to shifts in policies and practices, "Big Tech" companies are now more likely to remove or decrease the reach of policy-violating content and accounts. In the aftermath of the Charlottesville Unite the Right rally, Facebook removed several prominent far-right pages and Twitter unverified some prominent far-right actors (Donovan et al., 2019). More recently, "Big Tech" companies removed significant amounts of problematic content, including content that promotes the QAnon conspiracy theory (Conger, 2020); pushes the "Big Lie" alleging fraud in the 2020 United States presidential election (Collins & Zadrozny, 2021; Isaac, 2021); or spreads Covid-19 misinformation (Gold, 2021). This era of increased moderation also includes

content takedowns aimed at major world leaders, including the suspension of former President Donald Trump's accounts after the January 6 insurrection and the removal of Covid-19 disinformation from Brazilian President Jair Bolsonaro (Satariano, 2021; Twitter, 2021).

In one part a response to "Big Tech" companies taking more aggressive stances on content moderation, a growing group of "alt-tech" platforms have gained traction, often with users who have been de-platformed from mainstream, "Big Tech" platforms (Freelon et al., 2020, p. 3).[3] Far-right groups have long used online platforms for organising, but the shift away from "Big Tech" platforms brought increased attention to the "alt-tech" ecosystem (Donovan et al., 2019). These "alt-tech" platforms have attempted to market themselves as unfettered, unmoderated areas which prioritise unlimited free speech. This, they claim, is in contrast to "Big Tech" platforms, which they portray as censorious and at odds with principles of liberty and freedom.

3. Methods

3.1 Platforms observed

Prior to engaging in analysis of the platforms presented in this paper (Twitter, YouTube, Parler, and BitChute), we analysed policies developed and implemented by Facebook and Gab, another pairing of similar "Big Tech" / "alt-tech" social media platforms, to establish (1) the usability of our model; and (2) inter-coder reliability. This calibrating analysis was not included in our findings. Our research centres on a qualitative content analysis of platform policy (Hsieh & Shannon, 2005; Mayring, 2014). Content analyses are a common approach to research on platform policy, and have been used to study variations in policies in general (Jiang et. al., 2020) as well as specifically focusing on harassment policies (Pater et. al., 2016). The policies, which varied in length, location, and depth, were all compared against the same dimensions in our coding scheme.

3.2 Coding Scheme

We defined our codebook in three sections, which we describe as Content Categories, Enforcement Options, and Detection Mechanisms. Content Categories refer to codes which represent whether a particular platform's policy addresses content of that particular variety (like hate speech or discrimination). Enforcement Options refers to codes which describe how a platform can respond when content that goes against their policy is posted to their platform. Detection Mechanisms refer to codes which describe

how a platform surfaces and observes content that might violate its policies around objectionable content.

For each code and platform, we had the option of three values through which the platform could be described: yes, no, and unclear. Yes means that there is evidence in the platform's user-generated content policies which indicates that the platform restricts this kind of content, can act against offending content in this manner, or can detect content through this mechanism, according to a reasonable interpretation of the policy. No means that there is no evidence in the platform's policies which is reasonably interpreted to mean that the platform polices this form of content, takes this form of enforcement action, or detects offending content in this manner—note, this doesn't necessarily mean that the platform doesn't take this action, only that their publicly-visible policies do not reference it. Finally, Unclear means that there are reasonable interpretations of the platform's content policies for either yes or no, or for other reasons does not fit this binary classification well.

3.3 Limitations

Our work is limited in several ways which deserve acknowledgement. First, we studied four platforms; there are numerous platforms in existence, both "Big Tech" and "alt-tech," that remained unexamined. Therefore, it is possible other permutations reveal underpolicing or underrecognition of "alt-tech" behaviour on the whole, thus bringing it in line with digital marketplaces' laissez-faire approach to addressing problematic content on "Big Tech" platforms. Second, we did not address "alt-tech" platforms developed outside of a political, ideological purpose. Alternative social media is available for many different interest groups, of which "free speech" is just one, and focusing on the "alt-tech" scene aligned with the ideological right is not generalizable to other alternative social media. Third, it is imperative to note an immutable characteristic of platform policy research: change over time. From the date of authorship to the date of publication, it is entirely possible that "alt-tech" platforms bolster their policies or that "Big Tech" platforms reduce theirs. As such, it is possible changes could undermine the immediate value of this commentary. We are further limited by scarce data available about platforms' enforcement decisions (especially regarding how often and against whom). It is difficult for researchers across the board to obtain systemic data describing Twitter and YouTube's enforcement environments; it is virtually impossible to do the same for Parler and BitChute. This is likewise an issue for digital marketplaces, as the rationales for, and even times when, applications have been taken down from marketplaces are not publicly viewable. Finally, our work

is limited in scope to only being within a Western political context, and predominantly focused on the United States, with some influence from the United Kingdom (since BitChute is based in the U.K.). Lessons learned from these platform policies ought not to be generalised to alternative tech platforms in other political and cultural contexts.

3.4 Ethical Considerations

We feel the need to explain the ethical implications of this project. While we acknowledge there are benefits to content moderation in some scenarios, such as removing hate speech, there are also concerns such practices present. For example, content moderation has been shown to perpetuate systems of sexism (Gerrard & Thornham, 2020). Such biases are also prevalent when these systems are automated, as these systems perpetuate the viewpoints of their trainers, which represent a non-representative subset of viewpoints on the contested norms of what is appropriate content for the platform (Binns et al., 2017; Gillespie, 2020). These practices are highly dependent on who manages moderation, and who builds systems managing moderation, so simply advocating for more moderation will not necessarily improve digital environments, and may in fact harm historically marginalised communities (Birhane, 2021). Harms are particularly evident in cases when governments can compel platforms to moderate content, which can be used by authoritarian regimes to repress protests (Douek, 2021). Additionally, there are legitimate cases when keeping harmful content accessible, at least to some categories of observers, is necessary, such as when investigating human rights violations (Banchik, 2021). Furthermore, content moderation is an exploitative job with significant costs to its workers; it unacceptably impacts their health and wellbeing long-term (Roberts, 2019). Therefore, in terms of improving content policies, and content moderation as a whole, researchers and practitioners ought to consider how to make these processes more just and equitable (Gerrard, 2020).

4. Results

Below we include a table summarising our coding results for all four platforms.

Category	Twitter	Parler	YouTube	BitChute
Content				
Election-related	Yes	No	Yes	No
Health-related	Yes	No	Yes	No
Violence	Yes	Yes	Yes	Yes
Incitement to violence	Yes	Yes	Yes	Yes
Discrimination	No	No	No	No
Hate Speech	Yes	No	Yes	No
Fraud	Yes	Yes	Yes	Yes
Manipulated Media	Yes	No	Yes	No
Inauthentic Behavior	Yes	Yes	Yes	Yes
Hack-leak/hack-forge-leak	Yes	No	Yes	No
General Misinfo	No	No	No	No
General Disinfo	No	No	No	No
Enforcement				
Account Suspension	Yes	No	Yes	Yes
Account removal	Yes	Yes	Yes	Yes
Post Removal	Yes	Yes	Yes	Yes
Suppression	Yes	Yes	Yes	Yes
Strike allotment	Yes	Yes	Yes	Yes
Labeling of posts	Yes	Yes	No	Yes
Labeling of accounts	Yes	No	No	No
Political exemption	Yes	No	No	No
Influencer Adjustment	No	No	No	No
Newsworthy Adjustment	Yes	No	Unclear	No
Appeals Mechanism	Yes	Yes	Yes	Yes
Detection				
AI moderation	No	Unclear	Yes	No
Content moderators - paid	Yes	Yes	Yes	Yes
Community Moderators	No	Yes	Unclear	No
User flagging/reporting of posts	Yes	Yes	Yes	Unclear
User flagging/reporting of accounts	Yes	Yes	Yes	Unclear

5. Discussion

To explain our results, we analyse several notable differences across the “Big Tech” / “alt-tech” divide: (1) hate speech and discrimination policies; (2) untrustworthy content; (3) exceptions to enforcement actions; and (4) policy transparency. Next, we discuss digital marketplaces’ interactions with “Big Tech” and “alt-tech” platforms, offering several possible motivations for differential treatment, including a focus on platform policies over content.

Overall, "'alt-tech" and "Big Tech" platforms maintain some similar policy components. Across the four platforms we studied, most contained the same core sections: content restrictions (fraud, inauthentic behaviour, violence, and incitement to violence); enforcement actions (account and post removals, suppression, strike allotments, and content labels), and methods of moderation (paid content moderators). Despite these similarities, however, differences in policy depth, accessibility, and clarity clearly divided "Big Tech" platforms from "alt-tech" platforms. Those differences, we believe, may be one indicator capable of explaining why digital marketplaces appear to differentiate between "Big Tech" and "alt-tech" products.

5.1 Hate Speech and Discrimination Policies

One significant difference between mainstream and alternative platforms is the presence of language policing hate speech. Generally, mainstream platforms name hate speech as a sanctionable offence; alternative platforms' policies do not contain provisions that address hate speech. For example, YouTube's policy clarifies that "[w]e remove content promoting violence or hatred against individuals or groups based on . . ." an enumerated list of identities (YouTube, n.d.-e). By contrast, Bitchute's policy does not appear to mention hate speech even in passing. The same divide holds true when comparing Twitter against Parler. Although this appears to paint the mainstream platforms in a relatively favourable light (perhaps signalling that mainstream platforms are less willing to tolerate abject hatred toward certain groups of people), the grey area between hate speech and discrimination complicates what otherwise tempts a clear conclusion.

Despite maintaining hate speech as a sanctionable offence, all four platforms lack substantive policies sanctioning mere discrimination on the basis of race, gender, sexual orientation, and other identity-based characteristics.[4] Although we do not argue that sanctioning discrimination is favourable (not all discrimination is sinister; for example, most people discriminate between cereal brands to purchase at the grocery store), the line between what constitutes hate speech and what qualifies as discrimination is muddy.

One recent example is YouTube's interaction with the popular "Louder with Crowder" channel. Steven Crowder, a far-right political commentator, is known for producing videos addressing inflammatory events and issues. Although Crowder claims his channel is presently monetised, meaning YouTube permits Crowder to run advertisements through its "AdSense" program, YouTube has previously struck and demonetised it.[5] The demonetisation and initial ban followed a series of incidents, the

most high-profile of which contained a racist "spoof" of George Floyd's murder (CrowderBits, 2021). Several months later, Crowder's platform was reinstated. Although Crowder continues to make anti-Semitic, racist, and misogynistic comments on almost every episode of his show, the channel apparently remains untouched by further policy-driven consequences.

A similar problem pervades Twitter. Here, as is the case with YouTube, it is difficult to square how Twitter distinguishes hate speech from discrimination; it permits some obscenities to remain but takes policy-driven action against others. Twitter's high-profile, varied attempts to limit former U.S. president Donald Trump's hate speech are illustrative. To some extent, his hateful lamentations were permissible. For example, one tweet following protests in Minneapolis after George Floyd's murder read: "...when the looting starts, the shooting starts" (@realDonaldTrump, 2020). The tweet remained available for two hours before Twitter took action, despite the fact that the tweet appeared to target majority-Black communities and directly promised violence. Eventually, the platform labelled the tweet and noted that although it violated policy (against violence, rather than hate speech, despite racially charged undertones), public interest warranted its continued availability (Sprunt, 2020).

At a certain point, however, the platform began to take drastic action. Following months of escalation by both Twitter and Trump, the platform banned the sitting President from using its products (Twitter, 2021). It appears that the January 6 insurrection, and the President's use of Twitter during that time, drove that result (Twitter, 2021). Trump's continued attempts to stoke the tensions at the Capitol, and the explicit presence of white supremacists on its premises, were likely additional reasons to move forward with the ban. Although his status as president likely contributed to the longevity of his presence on the site, accounts he regularly re-tweeted (which deliver similarly problematic content) remain operational. The line separating hate speech from discrimination, and thus what is sanctionable, remains unclear.

Generally speaking, American platforms will not place a content ban on discriminatory content. Whether or not the platform claims a "free speech" ideology, First Amendment principles (speaking without interference from a centralised, governing body) pervade platform moderation practices. Further, adding a content ban on discriminatory content would significantly increase the already overwhelming amount of content for moderators to review, increasing the already high costs of moderation to levels platforms would likely be unwilling to pay (Schoolov, 2021). In order to

moderate at that scale, platforms may turn to AI; however, that is currently ineffective, and presents significant moral complications (Gillespie, 2020)[6]. As such, the hate speech to discrimination continuum is likely to remain fuzzy and obscure; although mainstream platforms maintain the ability to sanction some content on the basis of hate, the line between hate and discrimination is opaque enough to permit "Big Tech" platforms to behave like alternative platforms when it suits their bottom line or corporate politics.

Including hate speech policy seems to be one (of many) factors that exempt mainstream platforms from pressure exerted by digital marketplaces. For example, Twitter and YouTube faced no public blowback from Apple's App Store or Google's Google Play Store in the aftermath of the attack on the U.S. Capitol. Although there is ample evidence to suggest that mainstream platforms played a role in permitting insurrectionists to organise (Mac et. al., 2021), the presence of policy suggesting the impropriety of such content may have been a legitimate protection against their removal from a digital marketplace.

5.2 Untrustworthy content

Another notable difference between mainstream platforms and alt-tech platforms we studied is their approach to addressing untrustworthy and manipulated forms of content. For example, both mainstream platforms we studied have policies addressing manipulated media, and information stemming from hacked materials that might fall under a hack-forge-leak descriptor (Twitter, n.d.-o; YouTube, n.d.-c). In contrast, the "alt-tech" platforms we examined had no such content restrictions. While none of the platforms we studied had generalised policies on misinformation or disinformation, both mainstream platforms do have current policies regarding misinformation for election-related and health-related content, which our studied "alt-tech" platforms also lack[7]. Noteworthy is how recently "Big Tech" platforms added these specifications—had this study been done earlier, these policies would not have existed (Election Integrity Partnership, 2020; Virality Project, 2021). Alt-tech platforms do have some restrictions on untrustworthy content, such as fraud and inauthentic behaviour.[8] All four platforms studied had policies for both of these categories of content. But these are not as extensive, nor do they cover as many categories of content of questionable veracity, as do the policies of mainstream platforms.

5.3 Exceptions to Enforcement Actions

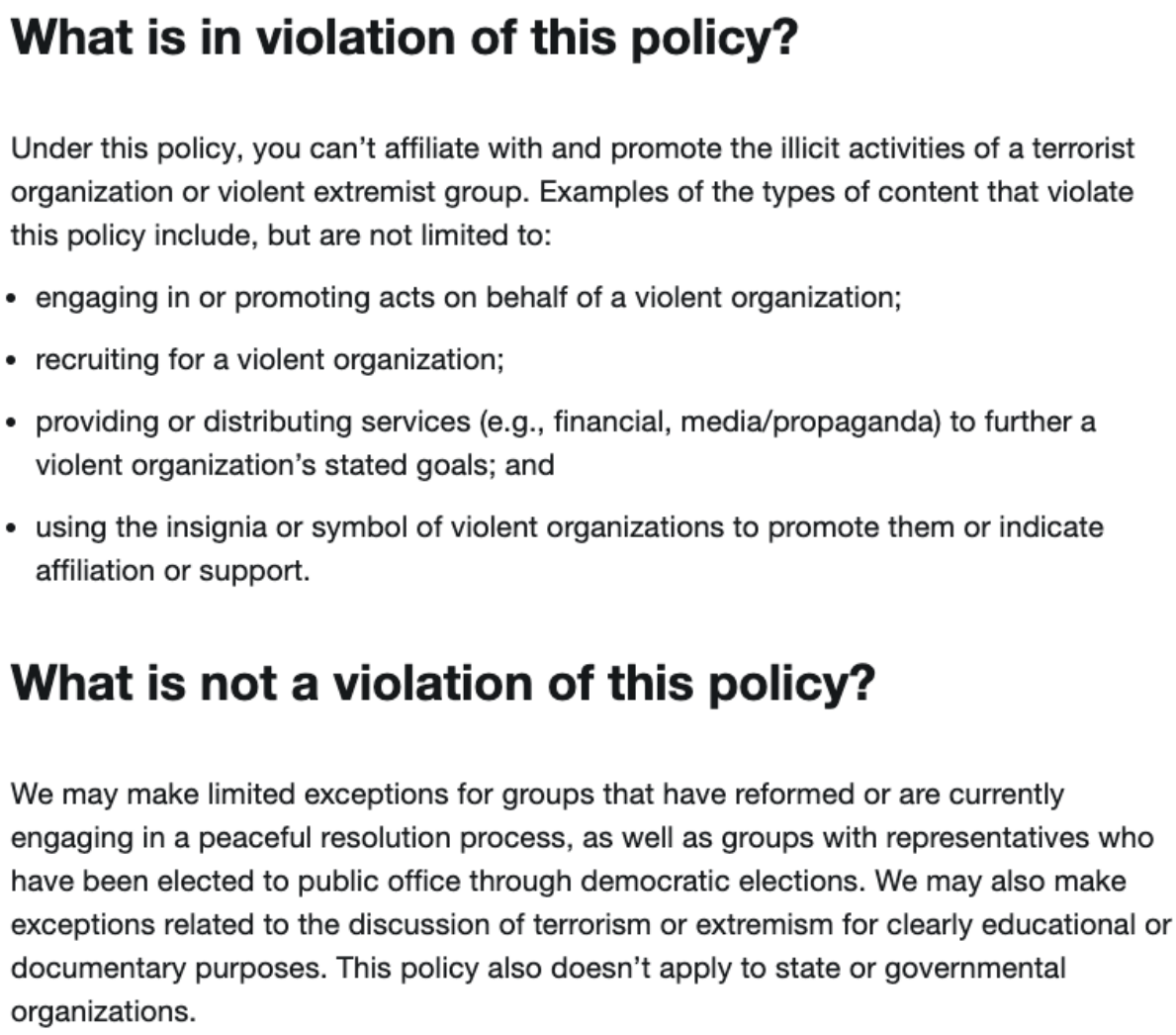
Many of the enforcement actions that alternative platforms allow themselves to take, at least from what is described in their policies, are similar to the actions taken by mainstream platforms. Other than Parler, which did not have a clear policy for temporary suspension of user accounts, all interventions we examined are described in policies on all four platforms. However, the mainstream platforms are more explicit in what exceptions they consider to their policies. For example, Twitter describes that they may allow otherwise-violating content to remain on their platform if it is in the “public interest” (Twitter, n.d.-g), and YouTube says that they make exceptions for content with "Educational, documentary, scientific, or artistic'' purposes (YouTube, n.d.-j).

Parler and BitChute describe no such exceptions. However, given the ideological, anti-censorship leanings of these platforms, this might be better described as viewing enforcement *as* the exception. Additionally, transparency on what exactly is enforced when it comes to content moderation, allowing us to see more definitively how common enforcement is, while somewhat limited on all platforms, is far more transparent on mainstream platforms than their alternative counterparts. Both Twitter and YouTube release transparency reports, which detail the amounts of content taken down (Twitter, n.d.-j; YouTube, n.d.-k). Twitter also makes limited sections of that data, such as those connected to information operations, accessible to researchers (Twitter, n.d.-d). In contrast, Parler and BitChute provide no such transparency data that we were able to find. While all platforms have given themselves room to interpret each decision, the rationales and justifications which the mainstream platforms use to make these decisions are often more readily apparent.

5.4 Policy Transparency and Clarity

Another example of differences across the “Big Tech”/ “alt-tech” divide is the accessibility and clarity of the documents which outline the platforms’ content policies. Because Twitter and YouTube have far more extensive guidelines, reading them in their entirety is more time-consuming. However, these platforms tend to be more explicit about what content violates their policies, and most forms of violative content are given their own specific policy documents. These also include clarifying examples on what does, or does not, qualify as policy violations. The screenshot below of Twitter’s violent organisations policy is one example. YouTube has a similar structure for many of their rules, and additionally has explanatory videos for several policies to

elaborate further, such as their child safety policy (YouTube, n.d.-a; YouTube Creators, 2019).



What is in violation of this policy?

Under this policy, you can't affiliate with and promote the illicit activities of a terrorist organization or violent extremist group. Examples of the types of content that violate this policy include, but are not limited to:

- engaging in or promoting acts on behalf of a violent organization;
- recruiting for a violent organization;
- providing or distributing services (e.g., financial, media/propaganda) to further a violent organization's stated goals; and
- using the insignia or symbol of violent organizations to promote them or indicate affiliation or support.

What is not a violation of this policy?

We may make limited exceptions for groups that have reformed or are currently engaging in a peaceful resolution process, as well as groups with representatives who have been elected to public office through democratic elections. We may also make exceptions related to the discussion of terrorism or extremism for clearly educational or documentary purposes. This policy also doesn't apply to state or governmental organizations.

Figure 2: An example Twitter policy that provides multiple clarifications of what kinds of content do and do not violate this particular policy (Twitter, n.d.-e).

In comparison, Parler's most analogous policy for Twitter's policy on violent organisations is their terrorism policy. The entirety of this policy is: "Terrorist organizations officially recognized as such by the United States are forbidden from using Parler, as is anyone—including state actors—recruiting for such organizations" (Parler, 2021c). What exactly is meant by the word "recruiting" in the prior quotation is not explained, so users may wonder if this would include using the symbols of an organisation, or providing financial services, as explained in the Twitter policy. Other

similar policies between Twitter and Parler demonstrate the same differences in elaboration and specificity.

While the policies for Parler and BitChute are spread across far fewer documents, however, this does not mean that these policies are more approachable and navigable. For example, Parler has only three content policy documents, but two of those documents were taken down and only accessible on the Internet Archive. Nonetheless, Parler's community guidelines maintained links to both of the missing policies (Parler, 2021b, 2021c). Twitter has similar issues, where documents were orphaned and not accessible to us through links in Twitter's policy documents, such as their policy on hacked materials (Twitter, n.d.-o). However, it is clear that while the "Big Tech" platforms' policies are far from easily navigable, having a smaller set of documents like the "alt-tech" platforms implement sacrifices specificity while not improving readability. Additionally, both Twitter and YouTube maintain policy search functions, allowing documents to be found and relevant sections to be quickly accessed. Parler and BitChute maintain no such functionality. As a result, when enforcement actions are taken by alt-tech platforms, they are likely to be even more inscrutable than the already difficult-to-parse moderation decisions of Big Tech regarding policy violations—the same decisions leading those on the ideological right to claim censorship and to grow the alt-tech ecosystem.

5.5 Engagement with and responses from digital marketplaces and intermediaries

The conversation about who regulates whom (and how) is ongoing and without clear answers. After the January 6 insurrection, however, it became clear that digital marketplaces are one site of private regulation capable of shifting platform behaviour. "Alt-tech" platform Parler was ousted from Apple's App Store, the Google Play Store, and Amazon's AWS servers. Although Parler has yet to return to AWS or the Google Play Store, Apple permitted Parler to re-enter its digital marketplace following an update to Parler's Terms of Service. During this same stretch of time, "Big Tech" platforms remained available for download across the same digital marketplaces, despite significant contributions to the mis- and disinformation environment during and after the 2020 election. This begs the question: to what degree, if at all, does policy shield platforms from regulation by digital marketplaces? What specific aspects of a platform's policy appear capable of warding off regulation from digital marketplaces? Although we do not arrive at a sure-fire answer to those questions, we

explore the possibility that robust platform policy matters a great deal to digital marketplaces - and the reasons for drawing that conclusion.

To begin, digital marketplaces are spaces where application developers and software users can interact with, purchase, or create products online. Applications downloaded from digital marketplaces are typically available on devices like laptops, smartphones, and tablets. Facebook, YouTube, and similar “Big Tech” platforms all maintain applications available for download on digital marketplaces. Some “alt-tech” platforms are also available for download. Within the last year, however, some “alt-tech” platforms’ unwillingness to moderate violent or hateful content drew punitive action from digital marketplaces and intermediaries. Following the January 6 insurrection on the United States Capitol, Apple indefinitely removed “alt-tech” platform Parler from its marketplace. In March, Apple then rejected Parler’s re-entry into its digital marketplace, citing a continuity of “highly objectionable content” on the application, including overwhelming hate speech and Nazi imagery (Lyons, 2021). Eventually, Apple reaccepted Parler into its App Store; Parler boasted that its automated content moderation could now better seek and remove the type of hateful content that Apple deemed unacceptable (Ghaffary, 2021). By contrast, Parler remains banned on Google’s Google Play store.

There are several potential reasons for different treatment “Big Tech” and “alt-tech” platforms receive; too, there is the very real possibility that digital marketplaces do express regulatory power or intend to regulate “Big Tech” platforms behind closed doors. At the time of initial submission, there existed no publicly available evidence that such conversations have transpired in the recent past. While recent revelations from leaked documents by Frances Haugen have shown that some such conversations have occurred between Facebook and intermediary platforms (Gambrell & Gomez, 2021; Scheck et al., 2021), the authors are as of final submission time unaware of public evidence of such conversations between the observed “Big Tech” platforms and intermediaries. Setting that possibility aside, it is plausible that pecuniary motivations limit digital marketplaces’ “regulatory” interactions with “Big Tech” platforms’ applications. To quantify the stakes, Twitter boasted 206 million daily active users in its 2021 Q2 shareholder report (Twitter Investor Relations, 2021). Also in early 2021, researchers reported that Parler hosted a usership of about 13 million (Thiel et al., 2021). To scale users in terms of downloads, “Big Tech” platforms YouTube and Twitter remain some of the highest-ranked applications by download on the App Store and the Google Play Store[9] (Apptopia, n.d.-b, n.d.-a). Parler and Bitchute do not appear on the enumerated “Top 50 Rankings Board” (Apptopia, n.d.-b, n.d.-a). Worries of blowback

from lucrative and in-demand "Big Tech" applications is a possible motivation for digital marketplaces to avoid taking action on the proliferation of misinformation and hate speech on popular social media platforms like YouTube,[10] Facebook, and Twitter.

Another potential reason for a more relaxed approach to moderating "Big Tech" platforms' applications could lie with the attitude of the product overall. Platforms like YouTube and Twitter are designed to host a variety of content, and neither were created with the specific intent to host *exclusively* ideological content. By contrast, platforms like Parler are designed specifically to host and promote a particular worldview in its content; that worldview has thus far prioritised nationalism, xenophobia, intense individualism, misogyny, anti-intellectualism, and racism (Freelon et al., 2020; C. Parker, 2018; C. S. Parker & Barreto, 2013). Because the latter are more likely to host a larger percentage of "objectionable content" overall, they are perhaps more likely subject to sanction. Although this is a plausible explanation for a difference in digital marketplaces' behaviour, it does not explain what is perhaps a willful ignorance of the fact that "Big Tech" platforms may still permit a higher *amount* of "objectionable content" due to the sheer size of their user bases (Thiel et al., 2021; Wagner, 2021).

Finally, it is worth considering what role policy plays in driving digital marketplaces' decision to regulate. On their faces, "Big Tech" platforms' policies are notably more robust than "alt-tech" platforms' policies. The former contains specific provisions about hate speech and incitement to violence, two key areas of concern digital marketplaces cited post-insurrection, whereas the latter often do not. "Big Tech" platforms provide navigable, accessible, and clear policy when juxtaposed with the usually stationery, limited, and confusing policies native to "alt-tech" platforms. Although robust policy alone is not evidence of *enforcement*, it seems to provide assurance that enforcement is possible—even likely. Because digital marketplaces have evidenced the importance of robust, specific policy regarding hate speech and violence in their interactions with Parler, it appears likely that language plays a highly important role in keeping platforms' applications available for purchase.

Although emphasising policy has its strengths, it provides a notable loophole for "Big Tech" platforms to evade punishment from digital marketplaces. Since policy is no guarantee of enforcement, this approach can be problematic. In the months since the 2020 election and the January 6 insurrection, misinformation and hate speech has flourished on "Big Tech" platforms regarding COVID-19 vaccination and a host of natural disasters. Although fraud is explicitly forbidden on "Big Tech" platforms,

related content is still available (Twitter, n.d.-i; YouTube, n.d.-f). Like the January 6 insurrection, policy-violating content is presently being under-regulated to the "real world" detriment of other platform users. In this manner, policy appears to be a shield from digital marketplace interference. The key difference between Apple's post-January 6 interruption of Parler's continuity in its marketplace and unaffected "Big Tech" platforms is the available policy backdrop. The same appears to be at work presently; anti-fraud policy is firmly in place across all platforms examined. This variety of treatment is a problem that deserves scrutiny. Because the United States Congress is unable to effectively act either due to gridlock or constitutional restrictions, digital marketplaces are one remaining source of possible platform regulation (Brannon, 2019). As long as actual enforcement appears to be lacking—especially on "Big Tech" platforms with significant user bases—digital marketplaces seem willing to recede from a regulatory role and instead collect on revenue shares.

6. Conclusion

Our research is interested in whether, and in what ways, content policies differ for social media platforms on opposite sides of the "Big Tech" / "alt-tech" divide. First, we found that there exist multiple key differences across the divide, such as content containing hate speech; untrustworthy content; exceptions to enforcement action; and overall content policy transparency. We also found several areas where policies were very similar, such as violent content; fraud; inauthentic behaviour; and the suite of enforcement actions which platforms make available to themselves. Additionally, we found differences in how intermediaries, such as digital marketplaces, interact with these social media platforms, which seem to be not directly related to the volume of objectionable content present on the platform.

Although there is poor transparency from all actors (platforms and marketplaces), enforcement actions and related topics are nevertheless important to study. Researchers should continue to explore platforms' content policies and reported enforcement actions as a vehicle to understand how platforms attempt to market themselves favourably to a variety of other actors—including but not limited to digital marketplaces (Gillespie, 2010). Further research into the area of content policy and moderation actions, such as platform and intermediary rationales for enforcement decisions, is needed. As politics becomes inextricably linked to social media, grasping whose motivations influence the regulation of online speech is critical. The question of who should participate in regulating online speech, and in what ways, will likely drive

research questions—and hopefully influence platforms for the better—in the years to come.

Acknowledgements

The authors would like to thank the University of Washington Center for an Informed Public, The National Science Foundation (Award 2027792), the John S. and James L. Knight Foundation, and Craig Newmark Philanthropies in helping to support and fund this research. We would also like to thank Dr. Kate Starbird, Dr. Emma Spiro, Dr. Jevin West, Dr. Ryan Calo, Dr. Kolina Koltai, Dr. Rachel Moran, and other colleagues who guided this research and provided crucial feedback.

About the Authors

Nicole Hope Buckley is a third-year law student at the University of Washington. She plans to enter the legal profession in fall of 2022 with a focus on the intersection between regulatory law, sociotechnical systems, and the future of democratic elections. Orcid ID: 0000-0002-2826-4265

Joseph S. Schafer is a fourth-year undergraduate computer science and ethics student at the University of Washington. He hopes to pursue graduate school in information science in order to understand how misinformation takes advantage of recently developed sociotechnical systems, like social media, to influence our society. Orcid ID: 0000-0002-6921-2074

Conflicts of Interest

The authors do not declare any conflicts of interest.

References

Anti-Defamation League, Color of Change, Common Sense Media, Free Press, NAACP, Sleeping Giants, LULAC, Mozilla, & NHMC. (2020). *Stop Hate for Profit*. Stop Hate for Profit. https://www.stophateforprofit.org

Apptopia. (n.d.-a). *Top Android Apps*. Retrieved July 25, 2021, from https://apptopia.com/store-insights/top-charts/google-play/overall/united-states?date=2021-07-01

Apptopia. (n.d.-b). *Top iOS Apps*. Retrieved July 25, 2021, from https://apptopia.com/store-insights/top-charts/itunes-connect/top-overall/united-states

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 20:1-20:27. https://doi.org/10.1145/3274289

Banchik, A. V. (2021). Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. *New Media & Society*, *23*(6), 1527–1544. https://doi.org/10.1177/1461444820912724

Bazelon, E. (2020, October 13). The Problem of Free Speech in an Age of Disinformation. *The New York Times*. https://www.nytimes.com/2020/10/13/magazine/free-speech.html

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. *ArXiv:1707.01477 [Cs]*, *10540*, 405–415. https://doi.org/10.1007/978-3-319-67256-4_32

Birhane, A. (2021). The Impossibility of Automating Ambiguity. *Artificial Life*, *27*(1), 44–61. https://doi.org/10.1162/artl_a_00336

BitChute. (n.d.-a). *Community Guidelines*. Retrieved July 25, 2021, from https://support.bitchute.com/policy/guidelines/

BitChute. (n.d.-b). *Content Moderation Policy*. Retrieved July 25, 2021, from https://support.bitchute.com/policy/content-moderation

BitChute. (n.d.-c). *Incitement to Hatred*. Retrieved July 26, 2021, from https://support.bitchute.com/policy-explanations/incitement-to-hatred

Brannon, V. C. (2019). *Free Speech and the Regulation of Social Media Content*.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

Center for an Informed Public, Digital Forensic Research Lab, Graphika, & Stanford Internet Observatory. (2021). *The Long Fuse: Misinformation and the 2020 Election*. https://purl.stanford.edu/tr171zs0069

Citron, D. K., & Wittes, B. (2017). The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity. *FORDHAM LAW REVIEW*, *86*, 24.

Collins, B., & Zadrozny, B. (2021, January 8). *Twitter bans Michael Flynn, Sidney Powell in QAnon account purge*. NBC News. https://www.nbcnews.com/tech/tech-news/twitter-bans-michael-flynn-sidney-powell-qanon-account-purge-n1253550

CrowderBits. (2021, April 7). *LIVE: Crowder Recreates George Floyd Arrest... | Louder With Crowder*. https://www.youtube.com/watch?v=O88knFix2fo

De Vynck, G., & Lerman, R. (2021, July 22). Facebook and YouTube are still full of covid misinformation. *The Washington Post*. https://www.washingtonpost.com/technology/2021/07/22/facebook-youtube-vaccine-misinformation/

Donovan, J., Lewis, B., & Friedberg, B. (2019). *Parallel Ports: Sociotechnical Change from the Alt-Right to Alt-Tech*. 49–65. https://doi.org/10.14361/9783839446706-004

douek, evelyn. (2021, June 2). More Content Moderation Is Not Always Better. *Wired*. https://www.wired.com/story/more-content-moderation-not-always-better/

Election Integrity Partnership. (2020, October 28). *Evaluating Platform Election-Related Speech Policies*. Election Integrity Partnership. https://www.eipartnership.net/policy-analysis/platform-policies

Freelon, D., Marwick, A., & Kreiss, D. (2020). False equivalencies: Online activism from left to right. *Science*, *369*(6508), 1197–1201. https://doi.org/10.1126/science.abb2428

Frenkel, S. (2021, July 19). White House Dispute Exposes Facebook Blind Spot on Misinformation. *The New York Times*. https://www.nytimes.com/2021/07/19/technology/facebook-misinformation-blind-spot.html

Gambrell, J., & Gomez, J. (2021, October 25). *Apple once threatened Facebook ban over Mideast maid abuse*. AP NEWS. https://apnews.com/article/the-facebook-papers-maid-abuse-94909f43c725af09522704348e35bd25

Gerrard, Y. (2020). Social media content moderation: Six opportunities for feminist intervention. Feminist Media Studies, 20(5), 748–751. https://doi.org/10.1080/14680777.2020.1783807

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, *22*(7), 1266–1286. https://doi.org/10.1177/1461444820912540

Ghaffary, S. (2021, May 17). *Parler is back in Apple's App Store, with a promise to crack down on hate speech*. Vox. https://www.vox.com/recode/2021/5/17/22441143/parler-apple-app-store-hate-speech

Gibson, A. (2019). Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society*, *5*(1), 2056305119832588. https://doi.org/10.1177/2056305119832588

Gillespie, T. (2010). The politics of 'platforms.' *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 2053951720943234. https://doi.org/10.1177/2053951720943234

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A., & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4). https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy

Gold, A. (2021, March 11). *Exclusive: YouTube removed 30,000 videos with COVID misinformation*. Axios. https://www.axios.com/youtube-removed-30000-covid19-vaccine-videos-misinformation-a8968086-95a4-4d5e-86da-0e22ddbc1b6a.html

Heilweil, R., & Ghaffary, S. (2021, January 8). *How Trump's internet built and broadcast the Capitol insurrection*. Vox. https://www.vox.com/recode/22221285/trump-online-capitol-riot-far-right-parler-twitter-facebook

Hill, M., & Freelon, D. (2020, September 5). *The threat of disinformation looms over the elections*. PBS NewsHour. https://www.pbs.org/newshour/show/the-threat-of-disinformation-looms-over-the-elections

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Isaac, M. (2021, January 11). Facebook will remove 'Stop the Steal' misinformation. *The New York Times*. https://www.nytimes.com/2021/01/11/us/facebook-stop-the-steal.html

Jackson, S. J., Bailey, M., & Foucault Welles, B. (2018). #GirlsLikeUs: Trans advocacy and community building online. *New Media & Society*, *20*(5), 1868–1888. https://doi.org/10.1177/1461444817709276

Jiang, J. "Aaron," Middler, S., Brubaker, J. R., & Fiesler, C. (2020). Characterizing Community Guidelines on Social Media Platforms. *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 287–291. https://doi.org/10.1145/3406865.3418312

Klonick, K. (2020). *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression* (SSRN Scholarly Paper ID 3639234). Social Science Research Network. https://papers.ssrn.com/abstract=3639234

Koltai, K. (2020). Vaccine Information Seeking and Sharing: How Private Facebook Groups Contributed to the Anti-vaccine Movement Online. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2020i0.11252

Le Gouvernement francais. (2018). *Against information manipulation*. Gouvernement.Fr. https://www.gouvernement.fr/en/against-information-manipulation

Lyons, K. (2021, May 17). *Parler returns to Apple App Store with some content excluded*. The Verge. https://www.theverge.com/2021/5/17/22440005/parler-apple-app-store-return-amazon-google-capitol

Mac, R., Silverman, C., & Lytvynenko, J. (2021, April 26). *Facebook Stopped Employees From Reading An Internal Report About Its Role In The Insurrection. You Can Read It Here.* BuzzFeed News. https://www.buzzfeednews.com/article/ryanmac/full-facebook-stop-the-steal-internal-report

Mayring, P. (2014). *Qualitative Content Analysis*.

National Center for Missing and Exploited Children. (n.d.). *Child Sexual Abuse Material*. Retrieved July 28, 2021, from https://www.missingkids.org/theissues/csam

Noble, S. U. (2018). *Algorithms of Oppression*. New York University Press. https://nyupress.org/9781479837243/algorithms-of-oppression

Papakyriakopoulos, O., Serrano, J. C. M., & Hegelich, S. (2020). The spread of COVID-19 conspiracy theories on social media and the effect of content moderation. *Harvard Kennedy School Misinformation Review*, *1*(3). https://doi.org/10.37016/mr-2020-034

Parker, C. (2018). The Radical Right in the United States of America. In J. Rydgren (Ed.), *The Oxford Handbook of the Radical Right* (pp. 630–649). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190274559.013.31

Parker, C. S., & Barreto, M. A. (2013). *Change They Can't Believe In*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691151830/change-they-cant-believe-in

Parler. (2021a, February 14). *Community Guidelines*. https://legal.parler.com/documents/guidelines.pdf

Parler. (2021b, March 3). *Guidelines Enforcement Process*. https://web.archive.org/web/20210303072739/https://legal.parler.com/documents/Guidelines-Enforcement-Process.pdf

Parler. (2021c, May 12). *Elaboration on Guidelines*. https://web.archive.org/web/20210512004242/https://legal.parler.com/documents/Elaboration-on-Guidelines.pdf

Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. *Proceedings of the 19th International Conference on Supporting Group Work*, 369–374. https://doi.org/10.1145/2957276.2957297

Pfaffenberger, B. (1992). Technological Dramas. *Science, Technology, & Human Values*, *17*(3), 282–312.

@realDonaldTrump. (2020, May 29). ....*These THUGS are dishonoring the memory of George Floyd, and I won't let that happen. Just spoke to Governor Tim Walz and told him that the Military is with him all the way. Any difficulty and we will assume control but, when the looting starts, the shooting starts. Thank you!* https://web.archive.org/web/20200529045316/https://twitter.com/realDonaldTrump/status/1266231100780744704

Roberts, S. (2016). Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Studies Publications*. https://ir.lib.uwo.ca/commpub/12

> Roberts, S. (2019). *Behind The Screen: Content Moderation in the Shadows of Social Media.* Yale University Press.

> Samek, G. (2018, February 2). *Greater transparency for users around news broadcasters*. Blog.Youtube. https://blog.youtube/news-and-events/greater-transparency-for-users-around/

Satariano, A. (2021, July 22). YouTube pulls videos by Bolsonaro for spreading misinformation on the virus. *The New York Times*. https://www.nytimes.com/2021/07/22/world/youtube-bolsonaro-covid.html

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*. The Wall Street Journal. https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953

Schoolov, K. (2021, February 27). *Why content moderation costs billions and is so tricky for Facebook, Twitter, YouTube and others*. CNBC. https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html

Shepardson, D. (2021, January 21). U.S. panel asks FBI to review role of Parler in Jan. 6 Capitol attack. *Reuters*. https://www.reuters.com/article/us-usa-trump-parler-idUSKBN29Q2FS

Sprunt, B. (2020, May 29). The History Behind "When The Looting Starts, The Shooting Starts." *NPR*. https://www.npr.org/2020/05/29/864818368/the-history-behind-when-the-looting-starts-the-shooting-starts

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). *Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing*. https://doi.org/10.9776/14308

Stegbauer, A. (2007). The Ban of Right-Wing Extremist Symbols According to Section 86a of the German Criminal Code. *German Law Journal*, *8*(2), 173–184. https://doi.org/10.1017/S2071832200005496

Thiel, D., DiResta, R., Grossman, S., & Cryst, E. (2021, January 28). *Parler's First 13 Million Users*. https://fsi.stanford.edu/news/sio-parler-contours

T-Mobile Support. (n.d.). *Pre-installed apps: Samsung Galaxy S8*. T-Mobile Support. Retrieved July 25, 2021, from https://www.t-mobile.com/support/devices/android/samsung-galaxy-s8/pre-installed-apps-samsung-galaxy-s8

Twitter. (n.d.-a). *Anti-Discriminatory Targeting Policy*. Retrieved July 26, 2021, from https://business.twitter.com/en/help/ads-policies/campaign-considerations/anti-discriminatory-targeting-policy.html

Twitter. (n.d.-b). *COVID-19 misleading information policy*. Retrieved July 22, 2021, from https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

Twitter. (n.d.-c). *Government and state-affiliated media account labels*. Retrieved July 23, 2021, from https://help.twitter.com/en/rules-and-policies/state-affiliated

Twitter. (n.d.-d). *Information Operations*. Twitter Transparency Center. Retrieved July 27, 2021, from https://transparency.twitter.com/en/reports/information-operations.html

Twitter. (n.d.-e). *Our policy on violent organizations*. Retrieved July 24, 2021, from https://help.twitter.com/en/rules-and-policies/violent-groups

Twitter. (n.d.-f). *Our range of enforcement options for violations*. Retrieved July 23, 2021, from https://help.twitter.com/en/rules-and-policies/enforcement-options

Twitter. (n.d.-g). *Public-interest exceptions to enforcement of Twitter rules*. Retrieved July 26, 2021, from https://help.twitter.com/en/rules-and-policies/public-interest

Twitter. (n.d.-h). *Reporting false information on Twitter for France*. Retrieved July 25, 2021, from https://help.twitter.com/en/rules-and-policies/france-false-information

Twitter. (n.d.-i). *The Twitter rules: Safety, privacy, authenticity, and more*. Retrieved July 22, 2021, from https://help.twitter.com/en/rules-and-policies/twitter-rules

Twitter. (n.d.-j). *Twitter Transparency Center*. Retrieved July 27, 2021, from https://transparency.twitter.com/en.html

Twitter. (n.d.-k). *Twitter's civic integrity and election fraud policy*. Retrieved July 22, 2021, from https://help.twitter.com/en/rules-and-policies/election-integrity-policy

Twitter. (n.d.-l). *Twitter's enforcement philosophy & approach to policy development*. Retrieved July 23, 2021, from https://help.twitter.com/en/rules-and-policies/enforcement-philosophy

Twitter. (n.d.-m). *Twitter's impersonation policy*. Retrieved July 22, 2021, from https://help.twitter.com/en/rules-and-policies/twitter-impersonation-policy

Twitter. (n.d.-n). *Twitter's policy on hateful conduct*. Retrieved July 22, 2021, from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Twitter. (n.d.-o). *Twitter's policy on the distribution of hacked materials*. Retrieved July 22, 2021, from https://help.twitter.com/en/rules-and-policies/hacked-materials

Twitter. (n.d.-p). *Twitter's sensitive media policy*. Retrieved July 22, 2021, from https://help.twitter.com/en/rules-and-policies/media-policy

Twitter. (n.d.-q). *Twitter's violent threats policy*. Retrieved July 22, 2021, from https://help.twitter.com/en/rules-and-policies/violent-threats-glorification

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

Twitter Investor Relations. (2021). *Q2 2021 Letter to Shareholders*. https://s22.q4cdn.com/826641620/files/doc_financials/2021/q2/Q2'21-Shareholder-Letter.pdf

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1079–1088. https://doi.org/10.1145/1753326.1753486

Virality Project. (2021, February 11). *Evaluating COVID-19 Vaccine Policies on Social Media Platforms*. The Virality Project. https://www.viralityproject.org/policy-analysis/evaluating-covid-19-vaccine-policies-on-social-media-platforms

Vis, F., Faulkner, S., Noble, S. U., & Guy, H. (2019). When Twitter Got #woke: Black Lives Matter, DeRay McKesson, Twitter, and the Appropriation of the Aesthetics of Protest. In A. McGarry, I. Erhart, H. Eslen-Ziya, O. Jenzen, & U. Korkut (Eds.), *The Aesthetics of Global Protest: Visual Culture and Communication* (pp. 247–266). Amsterdam University Press. https://doi.org/10.5117/9789463724913

Wagner, K. (2021, February 9). Twitter Jumps Most in a Year After Sales Top Estimates. *Bloomberg.Com*. https://www.bloomberg.com/news/articles/2021-02-09/twitter-beats-revenue-estimates-warns-of-slowing-user-growth

YouTube. (n.d.-a). *Child safety policy*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/2801999?hl=en

YouTube. (n.d.-b). *COVID-19 medical misinformation policy*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/9891785?hl=en&ref_topic=9282436

YouTube. (n.d.-c). *Elections misinformation policies*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/10835034

YouTube. (n.d.-d). *Fake engagement policy*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/3399767?hl=en&ref_topic=9282365

YouTube. (n.d.-e). *Hate speech policy*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/2801939?hl=en

YouTube. (n.d.-f). *Impersonation policy*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/2801947?hl=en&ref_topic=9282365

YouTube. (n.d.-g). *Report inappropriate content*. Retrieved July 26, 2021, from https://support.google.com/youtube/answer/2802027#zippy=

YouTube. (n.d.-h). *Spam, deceptive practices, & scams policies*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/2801973?hl=en&ref_topic=9282365

YouTube. (n.d.-i). *Violent or graphic content policies*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/2802008?hl=en&ref_topic=9282436

YouTube. (n.d.-j). *YouTube Community Guidelines & Policies—How YouTube Works*. YouTube Community Guidelines & Policies - How YouTube Works. Retrieved July 25, 2021, from https://www.youtube.com/howyoutubeworks/policies/community-guidelines/

YouTube. (n.d.-k). *YouTube Community Guidelines enforcement*. Retrieved July 27, 2021, from https://transparencyreport.google.com/youtube-policy/removals?hl=en

YouTube. (n.d.-l). *YouTube Trusted Flagger program*. Retrieved July 25, 2021, from https://support.google.com/youtube/answer/7554338?hl=en

YouTube Creators. (2019, May 1). *Child Safety Policy: YouTube Community Guidelines*. https://www.youtube.com/watch?v=pHOH5SDK0pc&t=1s

Footnotes

1. The Election Integrity Partnership (EIP) represented a collaboration between four of the U.S.'s top research institutions: The University of Washington's Center for an

Informed Public, The Stanford Internet Observatory, Graphika, and The Atlantic Council's Digital Forensics Lab. The EIP identified and analysed election-related misinformation and disinformation, communicating its findings to government, research-based, and tech-industry stakeholders. For a more detailed summary of the Election Integrity Partnership, please refer to "The Long Fuse: Misinformation and the 2020 Election." ↩

2. "Big Tech" refers to mainstream social media platforms well-known to the general public; "alt-tech" refers to generally conservative-leaning platforms with less restrictive content moderation practices (Freelon et al., 2020, p. 3). ↩

3. Deplatforming is a colloquial term referring to a user's ejection from a mainstream social media platform. Often, users who are deplatformed expressed extreme views or posted extreme content that violated platform policy so significantly that the policy prescribed removal. In some contexts, "deplatforming" is used as alternative verbiage to describe a user who was "removed" for posting hate speech, targeted harassment, or otherwise damaging disinformation. ↩

4. Although Twitter maintains a policy forbidding financial discrimination in certain contexts, the policy language itself appears to dictate that to "discriminate," as used, is meant "to choose between" rather than "to act with prejudice" (Twitter, n.d.-n, n.d.-a). ↩

5. A strike, or being "struck," refers to a system of "strikes" used to warn a platform's users when they have violated the platform's policies, and, when appropriate, justify account suspensions and/or removals. YouTube uses a three-strike system (YouTube, n.d.-k). ↩

6. AI can be used to moderate content, but handles nuance poorly, fails to generalise across contexts, and often makes incorrect decisions (Gillespie, 2020). Additionally, AI classification systems in general, of which moderation would be an example, have also been shown to be sexist and racist across numerous contexts (Buolamwini & Gebru, 2018; Noble, 2018). ↩

7. Note, Twitter does have a policy regarding false information in France (Twitter, n.d.-h). However, since this doesn't exactly correspond to our working definitions of misinformation and disinformation, nor is it a policy applicable to Twitter content outside of one nation, for simplicity we have chosen to ignore this in our primary argument. ↩

8. For example, Bitchute considers inauthentic behaviour to cover “Manipulation of metrics such as views, likes and/or subscriptions” (BitChute, n.d.-a). ↩

9. Note, YouTube does not appear in the top 50 for the Google Play Store. However, on many Android devices, YouTube is a pre-installed app, so this is not indicative of a lack of reach for YouTube as it is for Parler and BitChute, since they are not pre-installed on all Android devices (T-Mobile Support, n.d.). ↩

10. It is worth noting that YouTube is a subsidiary of Google; as such, it is all the more unlikely that the Google Play Store would ban a Google product from its digital “shelves.” ↩