International Journal of Communication 14(2020), 2589–2610 1932–8036/20200005

# Assessing Digital Threats to Democracy, and Workable Solutions: A Review of the Recent Literature

KATHLEEN M. KUEHN
Victoria University of Wellington, New Zealand

LEON A. SALTER
Massey University, New Zealand

Concerns surrounding the threats that digital platforms pose to the functioning of Western liberal democracies have grown since the 2016 U.S. election. Yet despite a preponderance of academic work in this area, the precise nature of these threats, empirical solutions for their redress, and their relationship to the wider digital political economy remain undertheorized. This article addresses these gaps with a semisystematic literature review that identifies and defines four prominent threats—fake news, filter bubbles/echo chambers, online hate speech, and surveillance—and constructs a typology of "workable solutions" for combating these threats that highlights the tendency to silo technical, regulatory, or culturally embedded approaches.

*Keywords: digital democracy, fake news, filter bubbles, echo chambers, hate speech, surveillance, surveillance capitalism*

Global movements such as the Arab Spring, Los Indignados, and Occupy temporarily reignited earlier assertions about the Internet's capacity to facilitate a more democratic, egalitarian public sphere. Many celebrated the Internet's techno-social and communicative affordances for enabling activists to mobilize against injustice with unprecedented immediacy and ease (Bennett & Segerberg, 2013; Papacharissi & de Fatima Oliveira, 2012) and for creating new forms of horizontalist democratic decision making (Graeber, 2013; Sitrin & Azzellini, 2014). These assertions added to other claims about the Web's capacity to uphold democratic traditions and values, including the democratization of information publishing (Castells, 2013; Jenkins, 2006), increasing political engagement (Miller, 2016), and government transparency and accountability (Kim & Lee, 2012; Nielsen, 2017; Wu, Ma, & Yu, 2017).

This idealism now seems quaint in the wake of events such as the Snowden revelations, Brexit vote, Trump election, Cambridge Analytica scandal, and Christchurch mosque shootings, in which the communicative functions of privately owned online platforms were weaponized to undermine the very democratic processes they promised to enhance. As Shapiro (2018) notes, today's Internet-based platforms now run in a "feudal,"

Kathleen M. Kuehn: kathleen.kuehn@vuw.ac.nz
Leon A. Salter: l.a.salter@massey.ac.nz
Date submitted: 2019–08–04

"dictatorial" fashion that undermine the sovereignty of users who have little input regarding key decisions. Moreover, this continues as platforms are blamed by government leaders for amplifying "terrorist and violent extremist content" (Christchurch Call, 2019, p. 1) and as new legislation seeks to counter the effects of rising antagonism on democratic engagement (e.g., the UK's Online Safety laws; EU's Terrorist Content Regulation; Australia's Sharing of Abhorrent Violent Material Bill; and Germany's Network Enforcement Act). Even the U.S. government, historically ambivalent toward regulating tech giants (Flew, Martin, & Suzor, 2019), recently fined Facebook US$5 billion over the platform's massive privacy breaches and spread of Russian disinformation.

In Foucauldian terms, a new "episteme" (Foucault, 1991, p. 54) or "discursive formation" (Foucault, 1972, p. 43), may be taking shape, one associated with a "series of objects of knowledge" (p. 43) that are enabling/enabled by a shift in power relations. These objects are identified here as four distinct but interrelated phenomena discursively constructed as "contemporary threats" to the Web's democratic promise: (1) fake news; (2) filter bubbles and echo chambers; (3) hate speech; and (4) surveillance. Although these threats are widely discussed in academic and popular discourse, there is little understanding around their precise scale and scope, (inter)relationships, or how to combat them. With so much information in circulation, the state of empirical knowledge is often muddied by sheer volume, interdisciplinary silos, and the political economic agendas of competing interests (academics, platforms, regulators, activists, etc.).

In an effort to organize the current state of knowledge, this semisystematic review of the literature scopes the empirical research on redressing the threats named earlier, which have been selected for their dominance across scholarly and political discourse. After a description of methods, the article identifies, defines, and contextualizes these digital threats before reviewing what evidence exists for their redress. We organize proposed empirically backed "solutions" into a typology of *technical, regulatory,* and *culturally embedded* approaches, concluding with a call for their integration. In the midst of much speculation, research investment, ongoing discourse, and legislative moves, we hope this review offers a consolidation of existing knowledge that can help guide prodemocracy scholars, activists, and legislators in their shared pursuit of building a more democratic Internet and society for all.

## Methods

A semisystematic literature review is a methodological approach aimed at providing an interdisciplinary overview of conceptually broad topics. This type of analysis is useful for "detecting themes, theoretical perspectives, or common issues within a specific research discipline or methodology" (Snyder, 2019, p. 335) to map research fields, synthesize the current state of knowledge, or identify future directions for research on a given topic. A semisystematic review is thus an appropriate method for narrativizing the state of empirical research on a topic as conceptually broad and interdisciplinary as the threats that social/digital media pose for liberal democracies.

As part of a larger scoping project on this issue (Elliott, Berenston-Shaw, Kuehn, Salter, & Brownlie, 2019), data collection and analysis proceeded over 18 months through a complex, iterative process that began with a literature review conducted by the second author in July 2018 addressing two questions:

*RQ1: What are the key opportunities for improving democratic participation through digital media?*

*RQ2: What are the key threats to the achievement of those opportunities?*

Initial searches on Google Scholar and Massey University's library database collated academic articles and gray literature published since 2012 using the terms "social media" AND "democracy," "digital democracy," "Internet democracy," "online democracy," and "e-democracy." This returned 59 articles and reports after filtering for relevance. The most salient threats featured in this corpus were identified by topic, which informed subsequent database searches that increased the corpus to a total of 109 journal articles/reports.

Based on their primacy in the literature and relevance to the New Zealand context at the time, the research team condensed 11 initial threats to four: *fake news, digital filter bubbles/echo chambers, hate speech*, and *surveillance*.[1] Following this, the first author undertook further research on a third question:

*RQ3: What evidence exists in the high-quality published literature on what works to overcome those threats?*

Analysis began by reading through all articles sourced for the first part of the study and collating evidence-based claims presented for countering these threats. Of 109 articles, only a handful offered evidence supporting their recommendations. To fill these gaps, separate searches on Google Scholar and Victoria University of Wellington's library database were conducted for empirical, solutions-oriented studies published since 2012. Keywords included the four threats with subsearches "data," "social media," and "government AND/OR state" as relevancy filters for "surveillance," and "evidence" and "empirical" to isolate studies making tested claims. The author coded each abstract in the total corpus of 224 journal articles, reports, and books for tested solutions, which were then thematically organized into a typology of *technical, regulatory*, and/or *culturally embedded* approaches.

Overall, a dearth of empirically backed research exists, particularly for fake news and surveillance. Although there is no shortage of empirical research identifying these objects as threats, comparatively few studies investigate or test strategies to combat them. Semisystematic reviews cannot realistically capture all published work (Snyder, 2019), and we acknowledge that search term and database choices, cognitive capacity, research expertise, and available resources limited our findings. However, we believe that the corpus represents a solid cross-section of existing interdisciplinary work. We encourage future research to build on its foundation.

## Digital Threats to Democracy

### *Fake News*

Concerns about fake news in popular, political, and academic discourse increased massively after the 2016 U.S. presidential election (Farkas & Schou, 2018). Investigative news (e.g., Silverman, 2016), a U.S. Senate report (Mueller, 2019), and scholarly work (e.g., Guess, Nyhan, & Reifler, 2018; Vosoughi, Roy,

[1] This research does not suggest that the four selected threats are the most important challenges facing global democratic futures; rather, their selection was informed by their salience in the literature and New Zealand national discourse at the time.

& Aral, 2018) have linked the exploitation of social media's affordances to the spread of misleading information about Donald Trump and Hillary Clinton, influencing the election in some way (Ziegler, 2018).

Fake news encompasses a variety of antidemocratic practices. As based on the literature, fake news is defined as informational content characterized by (1) *disinformation,* or purposefully misleading information (Ghosh & Scott, 2018); (2) *political goals* designed to shape public opinion of political actors, groups, or issues (Guo, Rohde, & Wu, 2020); (3) *reactivity,* or the capacity to evoke a strong emotional response (Persily, 2017); (4) *spreadability*, or rapid diffusion within and across networks (Vosoughi et al., 2018); and (5) *personalized and targeted* messages gleaned from a user's behavioral data (Ghosh & Scott, 2018; Isaak & Hanna, 2018). The latter two characteristics arguably differentiate fake news from 20th-century propaganda, linking it firmly to the goals of "surveillance capitalism" (Zuboff, 2019): the exploitation of personal data for behavioral prediction, modification, and profit. For example, Cambridge Analytica's targeted political advertising ostensibly swayed political opinion by simply following the logics of contemporary digital advertising (Isaak & Hanna, 2018).

However, the precise nature, reach, and presumed effects of fake news are contested in the literature (Farkas & Schou, 2018; Guess et al., 2018). Its relatively small audience size presents conflicting evidence about fake news' political reach and power (Fletcher, Cornia, Graves, & Nielsen, 2018; Nelson & Taneja, 2018); others argue that "fake news" is more effectively used as a rhetorical weapon by political adversaries seeking to delegitimize their opponents' views than disinformation presented as news (Farkas & Schou, 2018). Yet, given its salience in contemporary discourse, we main that its inclusion here is warranted, especially as platforms and governments continuously seek solutions to fix it. Moreover, innovations in dark patterns, bots, and other automated techniques that entrap, deceive, or mislead Internet users are increasingly commonplace in the user design experience, enabling the capacity for political misinformation to become more effective and opaque (Dieter, 2015).

### ***Filter Bubbles/Echo Chambers***

Filter bubbles—the algorithmically driven personalization of Internet content and searches—are both a technological affordance and limitation of the contemporary Web. While targeted content enhances the online experience by delivering relevant results, many critics blame *mass customization* for increased polarization and intolerance (College of St. George, 2018; Deb, Donohue, & Glaisyer, 2017). Unlike mass media's reliance on consumer demographics, automated sorting processes enable online content providers unprecedented opportunities to build and micro-target content on the basis of psychographic profiles (Isaak & Hanna, 2018). For critics, personalized online ecosystems narrow the scope of information, opinions, and resources that users encounter (Pariser, 2011). Filter bubbles arguably pose detrimental effects on civic discourse and democracy by facilitating intellectual and ideological isolation or preventing a shared understanding of contemporary ideas and issues. Today, filter bubbles are so ubiquitous that they often work in the background to our daily lives, shaping the information we receive "imperceptibly and without consent" (College of St. George, 2018, p. 5). These practices are central to the profits of today's most powerful Internet companies.

Online echo chambers result from interactions between filter bubbles and people's tendency to seek out information that comfortably corresponds with what they already know (i.e., "confirmation bias"; Berentson-Shaw, 2018, p. 36). Echo chambers can act as an identity-securing protection from the epistemological and ontological uncertainties created by opinions that counter our worldviews (Ceron & Memoli, 2016; Lu & Yu, 2020). Political content often exploits this vulnerability, reinforcing confirmation bias and strong polarizing effects (Ceron & Memoli, 2016).

While social media are not solely to blame for online echo chambers (Beam, Hutchens, & Hmielowski, 2018), network research on Twitter (Guo et al., 2020; Himelboim, McCreery, & Smith, 2013) demonstrates how certain platform features (e.g., "Follow") encourage ideological homogeneity and thus opinion polarization (Himelboim et al., 2013). A case study of Reddit (Massanari, 2015) similarly found that its content-sorting algorithm, which prioritizes the most popular and recent posts, can encourage "toxic technocultures" (p. 330) that subordinate dissenting opinions on contemporary issues. These sorting mechanisms give the appearance that some views are more widely held than they are, legitimizing systematic harassment of marginalized groups or opinion holders.

Some scholars argue that moral panics surrounding filter bubbles and echo chambers are overstated and blame technology for human problems (Bruns, 2019). Empirical studies in political communication find that user choices, not algorithms, limit informational diversity (Fletcher et al., 2018; Möller, Trilling, Helberger, & van Es, 2018). Human variables that moderate echo chamber effects include information sharing practices (Zimmer, Scheibe, Stock, & Stock, 2019), network homogeneity (Allcott & Gentzkow, 2017), rootedness in beliefs (Nguyen & Vu, 2019), level of political interest, and diverse media choices (Dubois & Blank, 2018). Recommendation algorithms may be no more bubble-inducing than human editors (Möller et al., 2018); those reliant on social media as their primary political news source do not always exhibit more polarized political beliefs and attitudes than traditional media consumers (Nguyen & Vu, 2019). Moreover, studies confirming filter bubbles/echo chambers are often based on nongeneralizable, single-platform case studies, not consumption across a media-diverse environment (Dubois & Blank, 2018). Importantly, this does not deny the existence of filter bubbles/echo chambers, but claims to their effects must account for the complex techno-social dimensions undergirding today's information consumption practices.

### *Hate Speech*

Hateful expressions toward marginalized groups can inflict profound psychological effects on their victims while perpetuating inequalities (Edstrom, 2016; Gardiner et al., 2016; Massanari, 2015). Social media amplifies hate speech, where it is valorized by the attention-seeking logics of surveillance capitalism in the form of clicks, likes, and ad revenue for platforms and users alike (Jakubowicz, 2017).[2]

Hate speech targeting racial, ethnic, and religious minority groups has become increasingly coordinated through the rise of neo-fascist, White nationalist groups such as the alt-right (Klein, 2017; Nagle, 2017). This decentralized group of far-right ideological conservatives appropriates anti-statist and

[2] Indeed, the rise of online hate speech is not an outcome of surveillance capitalism, but a range of geopolitical, ideological, and other sociocultural factors too numerous and complex to adequately unpack here.

transgressive ethics of earlier progressive Internet movements (e.g., Anonymous; Occupy) and practices of Internet "geek" culture appealing to younger generations (Nagle, 2017). While these groups have long flocked to nonmainstream newsgroups, bulletin boards, and other alternative spaces, these expressions have entered more mainstream media platforms in recent years, particularly YouTube (Jakubowicz, 2017). The Data and Society Research Institute shows how an increasingly influential network of conservative YouTube commentators exploits platform affordances that link shocking content, visibility, and monetary profit to perpetuate far-right ideology (Lewis, 2018).

Online hate speech has traditionally taken the form of anonymous defamatory and harassing posts (Jakubowicz, 2017; Klein, 2017) now increasingly articulated by microcelebrities using real identities to achieve notoriety and network influence (Lewis, 2018; Phelan, 2019). Many of these aspiring "influencers" aim to attain hero status within White-supremacy subcultures by inflicting both online/offline symbolic and actual violence toward minority groups—for example, the recent Christchurch mosque attacker.

Research finds that online hate speech toward women and members of the LGBTQI community especially targets public figures, politicians, and those who shape public debates—for example, journalists (Braithwaite, 2016; Edstrom, 2016; Guo et al., 2020). Trolling, abuse, or harassment is typically directed toward essentializing gender characteristics rather than the victims' work, influence, or leadership position (Edstrom, 2016; Massanari & Chess, 2018). *The Guardian* newspaper found among its own comment threads that of the 10 journalists who attracted the most hateful comments, eight were women and two were Black men (Gardiner et al., 2016). Hate speech undermines democratic engagement by discouraging minorities from writing controversial stories or circulating ideas that might elicit special attention from opponents (Edstrom, 2016; Massanari & Chess, 2018).

When the public sphere is particularly polarized, social media users have been found to "express less disagreeing opinion and exercise more withdrawal behaviors" (Chen, 2018, p. 3928). Fear of social isolation can contribute to a "spiral of silence" among minority opinion holders (Noelle-Neumann, 1974) or undermine democratic engagement as affected groups retreat from political debates and institutions for safe locations "focus[ed] on building intracommunal bonding" (Jakubowicz et al., 2017, p. xi). These self-siloing practices further reinforce echo chambers and polarization detrimental to an inclusive public sphere (Ray, Brown, Fraistat, & Summers, 2017).

### ***Digital Surveillance***

Increased levels of digital surveillance have similar chilling effects. Digital surveillance refers to the systematic collection and analysis of digital data by organizational actors for the purposes of regulating or governing behavior (Andrejevic, 2019; Coleman, 2019). The Snowden and Cambridge Analytica scandals exemplified how both state and commercial surveillance practices overlap in their exploitation of Big Data—the "data exhaust" (Zuboff, 2019, p. 210) produced by everyday interactions. Drawing from similar data sets, state/commercial actors employ different strategies and techniques for different end-goals; these practices (and the power relations embedded within them) have significantly different implications and effects on democracy (Lyon, 2014).

Studies analyzing public responses to the 2013 Snowden revelations find a positive correlation between perceptions of government surveillance and self-censorship, particularly among journalists and writers for whom unfettered speech is vital to the maintenance of a healthy, critical public sphere (Holcomb, Mitchell, & Purcell, 2015). Chilling effects extend to racial and religious minorities disportionately targeted by the state, evidenced by research that finds heightened self-censoring among Muslims and Black activists (Stoycheff, 2016; Stoycheff, Liu, Xu, & Wibowo, 2019).

Platform owners claim that they are legally obliged to respond to government data requests (Mackaskill & Dance, 2013). However, the Cambridge Analytica scandal saw Facebook providing "unfettered and unauthorized access to personally identifiable information" (Isaak & Hanna, 2018, p. 56) of more than 87 million users to a private analytics firm to sway the 2016 U.S. presidential election and Brexit vote. In this case, commercial and state interests in data-based analytical marketing overlapped with significant implications for democratic processes. Digital surveillance thus includes, but exceeds, behavioral marketing when used by political actors for purposes of population management and control.

In its commercial form, digital surveillance is the economic backbone of the contemporary Web, driver of the attention economy and surveillance capitalism generally. Interest in user data is supported by platform business models, which rely on data extraction for capital reinvestment (Srnicek, 2017). Yet the lack of transparency around data practices highlights "major issues of freedom of expression, political participation, and democratic governance" (Gorwa, 2019, p. 855). Attention economy logics largely dictate information received, undermining rational debate and the context for making informed choices needed to sustain a vibrant, functioning democracy.

## Empirically Based Workable Solutions

Having identified empirical support for the claims that fake news, filter bubbles/echo chambers, hate speech, and surveillance pose some measurable threat (however overstated) to democratic processes, the following section reviews the empirically tested solutions for countering these threats. These solutions are presented as technological, regulatory, and culturally embedded approaches. Notably, the four threats are not equally represented across categories; few empirically tested solutions are offered for fake news and surveillance. This lack likely speaks to their global and institutional complexity and, for fake news, may even serve as evidence that the effects of its consumption are overblown. This typology highlights these research gaps to encourage more de-siloed, interdisciplinary, and integrative ways of thinking about what long-term solutions might look like.

### *Technological Solutions*

The technological fix is commonly invoked as a means to resolve the technological problems threatening democracy. Research on fake news and filter bubbles proposes creating and testing new algorithms that better identify "credible" (European Commission, 2018, p.14) search content, or filtering tools that provide Internet users with more agency over the type of content they encounter (Emanuelson, 2017).

Much research suggests that platform design and affordances can drastically reduce most of today's digital threats. Affordances are the perceived range of possible actions related to the features of any given platform (Bucher & Helmond, 2018; Hutchby, 2001). Although not always used in their intended ways, design decisions are nonetheless influential in directing and constraining online behaviors and actions.

For example, design choices can reduce filter bubbles/echo chambers by engaging users in more ideologically diverse communities. Several studies confirm that features (e.g., tools, interface) available on noncommercial platforms afford different modes of interaction and environments (e.g., deliberative, asynchronous) than commercial spaces (Ombler, Russell, & Rivera-Munoz, 2016; Stohl, Stohl, & Ganesh, 2018). When user identities are less curated toward a consumerist framework, less performative environments can yield more considered dialogue, discussion, and debate.

Research also confirms a positive association between design and civility. Two case studies analyzing local government's use of the noncommercial deliberation tool Loomio to debate proposed policy changes found that the platform successfully created more in-depth, inclusive, and less polarizing discussion than offline debates (Ombler et al., 2016; Stohl et al., 2018). One study found that Loomio's temporal and discursive affordances facilitated "a space where all views can be considered, and everyone can still be heard" (p. 23) without time-based anxieties and other restrictions (Ombler et al., 2016). The second study found Loomio fostered "an inclusive discussion" (Stohl et al., 2018, p. 246) that gave visibility to an important but controversial public issue that brought together previously marginalized voices. Unlike polarized mainstream contexts that push minority groups further into the margins, noncommercial, deliberative fora can broaden the public sphere by facilitating inclusivity and rational debate, and create safer spaces for marginalized groups. These affordances suggest that deliberation can move beyond debate to achieve a higher order of democratic engagement—for example, achieving rational consensus or collective agreement. These findings complement e-democracy research that finds a positive association between perceived government transparency and trust in local government (Kim & Lee, 2012; Nielsen, 2017).

Platform design has also been tested as a technological fix for reducing online hate speech, particularly around issues of online anonymity. Whereas early scholarship celebrated anonymity's liberating affordances (e.g., Baym, 2010), current empirical research correlates anonymity to increased incivility—for example, trolling generally hinges on anonymous users with "fake" accounts (Edstrom, 2016; Galán-García, de la Puerta, Gómez, Santos, & Bringas, 2017). Some research finds that prohibiting anonymous commenting minimizes abuse (Rowe, 2015; Santana, 2014). Forced preregistration has been shown to solicit qualitatively better, but quantitatively fewer, comments by requiring extra effort and/or enabling accountability (Bakker, 2010; Santana, 2014). Rowe (2015) found that anonymous comments were comparatively more uncivil and personally insulting than verified Facebook comments on the same news stories, concluding that visibility to one's wider network functioned as a sanctioning mechanism to enable more civil political discussions. However, another study found that anonymous news comments were of comparatively higher quality than nonanonymous comments on news agencies' corresponding Facebook pages, concluding, "Facebook will provide few comments, will kill the trolls, but will not result in making the conversation more interesting" (Hille & Bakker, 2014, p. 572). Identity verification systems might encourage more civil discourse in some contexts, but with fewer people contributing to the conversation. The efficacy

of identity-verification systems is still debated and requires testing across different contexts in combination with other solutions.

Platform design is also presented as a technical fix for combating online surveillance. Viable proposed solutions include engineering stronger privacy settings into digital platforms, privacy-enhancing software (e.g., ad blockers/trackers; data encryption software), or platform models that do not collect and exploit user data. These tools not only circumvent surveillance, but have also influenced public debate around invasive tracking practices (Flew et al., 2019; Fuchs & Trottier, 2017; Gehl, 2018; Narayanan & Reisman, 2017). Evidence supporting their efficacy, however, is often prescriptive, anecdotal, or informed by producers of these tools—perhaps treating the symptom rather than the cause.

Content moderation is another commonly cited technological fix. This solution overlaps with regulatory and culturally embedded solutions, but is categorized here for its technological base. Calls for enhanced content moderation policies at large intermediaries are gaining traction as a means of combating these four digital threats, but are based on limited empirical testing. The general consensus contends that content moderation processes that combine technical and social (human) responses are more effective than automation alone. User flagging and removal has been a demonstrably effective, albeit labor-intensive, way of removing offensive speech (Crawford & Gillespie, 2016; Pöyhtäri, 2014). Rising public pressure to enhance content moderation has incentivized some platforms to design more user-driven moderation tools; Twitter recently unrolled features enabling more content control via tools that hide or flag undesirable content, verify trusted accounts, and enhance content "quality" by filtering and disabling notifications (Klonick, 2015). Although these features enable users to manually or semiautomate content moderation, they also threaten to oversanitize online spaces, intensify filter bubbles/echo chambers, and disconnect users from wider network affordances.

Technical advances in semi- or fully automated systems, including deep learning, show increased promise in identifying inappropriate content while drastically reducing human labor (Binns, Veale, Van Kleek, & Shadbolt, 2017; Delort, Arunasalam, & Paris, 2011). A study on Reddit's 2015 use of an automatic keyword identification tool to sanction two hate speech-laden subreddits found that it effectively reduced hate speech by 80% and discontinued use of associated accounts (Chandrasekharan et al., 2017). Galán-García and colleagues (2017) tested machine-learning algorithms to successfully identify anonymous cyberbullies at a Spanish school. However, tracking techniques raise a number of concerns around algorithmic sorting and institutional surveillance, particularly in educational settings. Deidentification initiatives ostensibly create new threats to democratic rights while trying to combat others.

Other ongoing research aims to advance a more holistic approach that semiautomates content moderation via more transparent classification systems that account for context and offer explanations for content deletion (Risch & Krestel, 2018). Because empirical research finds text-based mining to be insufficient on its own, researchers are turning toward nontext features such as user characteristics as potential data sets for detecting incivility online. One such study combined certain user features with textual features to slightly improve the performance of automated classification results in hate speech detection models (Unsvåg & Gambäck, 2018). This tactic too, however, functions on the submission of users to more surveillance.

Civil discourse and other speech/engagement goals are unlikely achievable through moderation alone. Researchers increasingly acknowledge that neither automated nor manual classification systems are ever neutral or free from human bias (Binns et al., 2017; Gillespie, 2018). And while combining automation and human labor may be most efficacious, these efforts are continuously undermined by other challenges. Cultural and contextual differences, for example, pose a considerable conflict for global platforms to standardize content "appropriateness"; that is, what's appropriate in one context isn't always in another. Even still, the complex issues undergirding effective moderation have not tempered claims that "more moderation" is the answer to combating antidemocratic behaviors/practices.

### *Regulatory Solutions*

Regulatory solutions are another common, but generally untested, means for combating digital threats to democracy, particularly for redressing state/commercial surveillance. Such calls posit that government-enforced regulations around data and information management can best prevent breaches and abuse.

A case study of the Singaporean Data Protection Act 2012 finds some evidence that giving data protection authorities sufficient authority to prosecute misdemeanors can positively influence the data management practices of private intermediaries (Lanois, 2016). However, little evidence suggests that these changes will reduce surveillance or data collection as much as they will regulate how data are stored, accessed, and used. These findings support other expert opinion calling for regulatory changes to data privacy policies, although few question why data collection occurs in the first place (e.g., Flew et al., 2019; Internet Governance Forum, 2015). The efficacy of the EU's General Data Protection Regulation model is also widely debated, but its broader effects are currently unknown (see Kamara, 2017).[3]

On a broader scale, a growing but untested regulatory solution pervading academic and popular discourse is the call to make large intermediaries legally responsible for the information they publish (Flew et al., 2019; Marda & Milan, 2018). These debates range from breaking up Big Tech monopolies, to regulating social media platforms as publishers, media companies, or utilities, to enacting stronger data protection policies and oversight, etc. These solutions are applicable to all four digital threats and their effects are largely speculative, resting on the supposition that enforcing platforms to be socially responsible is a starting point for resolving these issues. Much work remains to be done in this area.

### *Culturally Embedded Solutions*

Culturally embedded solutions encompass a range of nontechnical, nonregulatory practices designed to empower Internet users with more agency and control over their online experiences. These involve people-centered, community-embedded solutions with a civic dimension: for example, education, community engagement, online/offline activism, and other sociocultural responses.

[3] See also Marda and Milan's (2018) critique of some legislative approaches taken by governments lacking in technical expertise.

Education is frequently positioned as a culturally embedded approach to these four threats, such as training in critical thinking and media/information literacy—especially around the function and power of algorithms (European Commission, 2018). This solution is commonly invoked in the literature on fake news, although we did not locate studies that have tested the efficacy of educational strategies' influence over the ability to discriminate fake news from authentic content. Alternatively, other research looks at the roles that civil society, advocacy organizations, and grassroots communities play in mitigating the effects of digital threats to democracy—that is, tactics that embody new modes of civic engagement.

Engaging civil society and advocacy organizations has been cited as a way of building knowledge communities and resilience to online disinformation, incivility, and filter bubbles, or in the circulation of countersurveillance techniques. Mobilizing new or existing support networks has been shown to develop fast, effective reporting mechanisms and support networks in combating online hate speech. Advocacy and civil society organizations such as All Together Now, the Federation of Ethnic Communities' Councils of Australia, and Australia's Online Hate Prevention Institute (OHPI) have demonstrated some success with online crowdsourcing tools that identify, track, report, and/or remove hate speech (Bodkin-Andrews, Newey, O'Rourke, & Craven, 2013; Jakubowicz et al., 2017; Oboler & Connelly, 2018; Sweet, Pearson, & Dudgeon, 2013) or highlight moderation gaps among larger online platforms (e.g., Online Hate Prevention Institute [OPHI], 2012).

In lieu of few alternatives, many marginalized groups form online support communities based on a shared sense of identity. Although these are not solutions for preventing hate speech, they suggest that a networked approach is perceived as a useful strategy for mitigating its effects. Case studies find that resilient communities such as @IndigenousX, a grassroots Twitter community of Indigenous Australians, function as support networks to victims of hateful attacks (Jakubowicz et al., 2017). Some groups effectively use Twitter as "participatory journalism" (Sweet et al., 2013, p. 104), building counternarratives to counteract racism directed at indigenous groups, while hashtag movements such as #BlackLivesMatter (Ray et al., 2017) provide a space to construct collective identities that validate victims' opinions and disclosures in an environment of reciprocity (Walton & Rice, 2013).

Coordinating diverse stakeholders to apply pressure to private intermediaries has had demonstrable effect in removing, but not preventing, hateful content online. Facebook eventually removed hateful content toward Aboriginal Australians after pressure from OPHI—not because it violated Facebook's Terms of Service (ToS), but because complaints spanned a broad and diverse range of actors across civil society, advocacy groups, regulators, and users. Sustained pressure from diverse stakeholders tends to garner mainstream media attention, threatens brand reputation of platforms, and can result in the forced removal or moderation of hateful content (Gagliardone, Gal, Alves, & Martinez, 2015; OHPI, 2012). However, many deem this process too slow. Removal speed is considered essential to diffusing the power of hate speech and trolling; the longer hateful content remains online, the more damage it inflicts on victims while empowering its perpetrators (OPHI, 2012). Early content removal may limit the scale of exposure, but does not limit or prevent hate speech itself.

The efficacy of culturally embedded solutions is less clear when it comes to surveillance. The opacity of state/commercial surveillance practices presents significant challenges to empirical researchers looking

to find and test solutions to their silencing or chilling effects. Opinion is divided over the influence that Snowden's revelations actually had over state Western surveillance powers. Snowden's lawyer Ben Wizner (2017) claims that public outrage and activism positively influenced surveillance discourse and policies, which effectively curtailed a range of surveillance programs and authority within the U.S. intelligence community and Big Tech. Others contend the "chorus of outrage by policy-makers, the media, civil society activists, and the general public" (Pohle & Van Audenhove, 2017, p. 1) had little effect on the West's intelligence-gathering practices; Hintz and Dencik (2016) found that initial Snowden debates in the UK press were critical of state surveillance, but ultimately supported the consolidation and expansion of state powers. For example, post-Snowden legislative changes such as the UK's Investigatory Powers Act (2016) and New Zealand's Intelligence and Security Bill (2017) added some transparency to previously opaque systems, but expanded state surveillance powers nonetheless.

## Discussion and Conclusion

This article identifies four prominent contemporary threats to digital democracy and presents a typology of solutions for their redress. Although these four threats combine to contribute to decreased trust in the Internet and our democratic institutions, few (if any) studies discuss them as mutually constituted phenomena or derived from the same structural conditions. They are rarely framed by empirical researchers as an outcome of a political economy based on data-driven capital accumulation (Fuchs, 2015; Zuboff, 2019), but as individual issues that can be fixed by technical and regulatory measures (or, rarely, culturally embedded approaches).

Recapturing the Web's most basic utopian promises first requires understanding what prevents their realization. Following scholars such as Srnicek (2017), Fuchs (2015), and Ghosh and Scott (2018), we contend that the inexorable rise of fake news, polarization, hate speech, and surveillance should be approached from a systemic perspective that acknowledges their interconnection as opposed to isolation, and intimate connection to data-driven capitalism (Srnicek, 2017; Zuboff, 2019). They are, together, phenomena exasperated by an economic system based on the exploitation of user data through the advertising practices of audience segmentation and behavioral profiling. The evidence of and opinion on the removal of hate speech, for instance, suggest that both light-touch voluntary ethical codes (Alkviadou, 2019) and more heavy-handed state legislation (Marda & Milan, 2018) are inadequate for addressing root causes or a political economy that profits from controversial material that "draws and holds consumer attention" (Ghosh & Scott, 2018, p. 4). As such, redressing these threats requires more than regulation, but a multifaceted, integrative approach.

Over the course of writing this review, the political climate has changed to one in which governments are rapidly looking to "do something" about the threats that digital platforms pose to democratic processes. Examples currently floated by Western governments range from breaking up Big Tech monopolies, global taxation plans, and new data sovereignty measures, to increasing transparency and accountability around data collection and privacy. Although a step in the right direction, their regulatory focus does little to challenge the structure of surveillance capitalism itself; for example, corporate divestment might increase competition, but it doesn't overturn the industry's primary business model of data exploitation or encourage digital literacies or the democratic governance of platforms more broadly.

Most current government proposals also preclude more holistic approaches that might simultaneously address some of the bigger geopolitical, cultural, and ideological issues motivating the Web's weaponization.

Moreover, large intermediaries anticipating regulation have advanced their own in-house solutions to address some of these threats—for example, Twitter's banning of political ads and Facebook's new algorithms that delete hate speech and fake news. Yet these self-regulatory "technical fixes" promise to resolve issues (e.g., filter bubbles, echo chambers, fake news) that a growing scholarly consensus finds widely overblown. These commitments by platforms seem little more than a discursive strategy designed to avoid stricter government regulations and motivated more by bottom-line concerns than actual harm done to users or democratic processes. Rather than challenge the political economy of surveillance capitalism, these self-regulatory fixes paradoxically ensure its reproduction while encouraging the belief that better technology can always resolve the very problems that technology (or surveillance capitalism) helps create.

A siloed approach is unlikely to resolve the broad economic influence and cultural power of social media platforms. Research might follow on Betkier's (2018) comparatively holistic approach for tempering social media's power by bridging the areas of law (state-imposed legislation), code (software-engineered solutions), norms (voluntary codes of ethics), and market forces (for example, removing the financial incentives to exploit user data). In this model, privacy mechanisms can be engineered into code and mandated by law, curbing the financial incentives of the attention economy and, therefore, the market (and cultural) power of the big platforms. Such integration closes the myriad gaps and loopholes that siloed approaches inevitably invite.

A siloed approach is also unlikely to resolve the broad economic and cultural power of social media platforms by regulating them as media or publishing companies. In the pre–social media era, global media conglomerates were subject to some media regulation across nation-states, but were still able to wield considerable ideological influence (Sinclair, 2017). Regulatory fixes alone will not create populations of critical or digitally literate thinkers with the requisite skills needed to navigate today's information economy or support victims of hate speech and incivility. Thus, community-embedded approaches (e.g., education, advocacy programs) that facilitate resourcing, support, and sites of identification still have an important role to play. Significantly more attention is needed here.

Culturally embedded approaches can inform the contextual and cultural complexities of regulating a globally networked social and informational environment. Internal content moderation policies that currently prioritize private and legal interests over social justice or advocacy-related goals (Roberts, 2016) can be reoriented toward the latter via multistakeholder initiatives that combine technical resources with the culturally embedded labor of users *and* experts, specialists, or community advocates trained in suicide prevention, human trafficking, child exploitation, domestic violence, terrorism, or other forms of harm. Content moderation must become an organizational priority that merges customer service, security, privacy, safety, marketing, branding, personnel, and the law to create a unified approach to resolving this complex issue. In other words, increasing accountability requires a matrix of technical, regulatory, and culturally embedded interventions that reduce harm and balance vibrant debate and civility without sanitizing the Web into an antidemocratic, authoritarian space.

Rescuing democratic rights from the contemporary political economy hinges on a large-scale social response. As surveillance capitalism evolves, new forms of collective, collaborative action that connects users/consumers with the market and state must be invented to prevent the total disconnection of economic production from politics and society. Zuboff's (2019) recommendations here map onto work being done by advocacy groups such as OHPI and autonomist Marxists, who engage multiple vested interests in their lobby for broader structural changes across the political economy and culture. We hope this overview and evaluation of the state of the literature inspires new and creative collaborations between relevant stakeholders working toward these ends.

## References

Alkiviadou, N. (2019). Hate speech on social media networks: Towards a regulatory framework? *Information & Communications Technology Law*, *28*(1), 19–35. doi:10.1080/13600834.2018.1494417

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. doi:10.1257/jep.31.2.211

Andrejevic, M. (2019). Automating surveillance. *Surveillance & Society, 17*(1/2), 7–13. doi:10.24908/ss.v17i1/2.12930

Bakker, P. (2010). Journalistiek zonder infrastructuur—een reële optie?[Journalism without infrastructure—a realistic option?]. *Tijdschrift voor Communicatiewetenschap*, *38*(3), 250–258.

Baym, N. K. (2010). *Personal connections in the digital age*. New York, NY: Polity.

Beam, M. A., Hutchens, M. J., & Hmielowski, J. D. (2018). Facebook news and (de)polarization: Reinforcing spirals in the 2016 U.S. election. *Information, Communication & Society*, *21*(7), 940–958. doi:10.1080/1369118X.2018.1444783

Bennett, W. L., & Segerberg, A. (2013). *The logic of connective action: Digital media and the personalization of contentious politics.* New York, NY: Cambridge University Press.

Berentson-Shaw, J. (2018). *A matter of fact: Talking truth in a post-truth world*. Wellington, New Zealand: Bridget Williams Books.

Betkier, M. (2018). *Moving beyond consent in data privacy law. An effective privacy management system for Internet services.* Wellington, New Zealand: Victoria University Press.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot?

Inheritance of bias in algorithmic content moderation. In G. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social informatics: Lecture notes in computer science* (pp. 405–415). Wiesbaden, Germany: Springer.

Bodkin-Andrews, G., Newey, K., O'Rourke, V., & Craven, R. (2013). Promoting resiliency to counter racism: The lived wisdom within Aboriginal voices. *InPsych: The Bulletin of the Australian Psychological Society Ltd*, *35*(4), 14–24.

Braithwaite, A. (2016). It's about ethics in games journalism? Gamergaters and geek masculinity. *Social Media + Society*, *2*(4), 1–10. https://doi.org/10.1177/2056305116672484

Bruns, A. (2019). *Are filter bubbles real?* Cambridge, UK: Polity.

Bucher, T., & Helmond, A. (2018). The affordances of social media platforms. In J. Burgess, A. Marwick, & J. Poell (Eds.), *The SAGE handbook of social media* (pp. 223–253). London, UK: SAGE Publications.

Castells, M. (2013). *Communication power* (2nd ed.). Oxford, UK: Oxford University Press.

Ceron, A., & Memoli, V. (2016). Flames and debates: Do social media affect satisfaction with democracy? *Social Indicators Research*, *126*(1), 225–240. doi:10.1007/s11205-015-0893-x

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (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*(CSCW), 1–22. https://dl.acm.org/toc/pacmhci/2017/1/CSCW

Chen, H. T. (2018). Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors. *New Media & Society*, *20*(10), 3917–3936. doi:10.1177/1461444818763384

Christchurch Call. (2019). *The Christchurch Call to action: To eliminate terrorist and violent extremist content online*. Retrieved from https://www.christchurchcall.com/christchurch-call.pdf

Coleman, G. (2019). How has the fight for anonymity and privacy advanced since Snowden's whistle-blowing? *Media, Culture & Society, 41*(4), 565–571. doi:10.1177/0163443719843867

College of St. George. (2018). *Democracy in a post-truth information age* (Background paper). Retrieved from https://www.stgeorgeshouse.org/wp-content/uploads/2017/10/Background-Paper.pdf

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint*. New Media & Society, 18*(3), 410–428. doi:10.1177/1461444814543163

Deb, A., Donohue, S., & Glaisyer, T. (2017). *Is social media a threat to democracy?* Retrieved from https://www.omidyargroup.com/wp-content/uploads/2017/10/Social-Media-and-Democracy-October-5-2017.pdf

Delort, J. Y., Arunasalam, B., & Paris, C. (2011). Automatic moderation of online discussion sites. *International Journal of Electronic Commerce*, *15*(3), 9–30. doi:10.2753/JEC1086-4415150302

Dieter, M. (2015). Dark patterns: Interface design, augmentation and crisis. In D. M. Berry & M. Dieter (Eds.), *Postdigital aesthetics* (pp. 163–178). London, UK: Palgrave Macmillan.

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society, 21*(5), 729–745. doi:10.1080/1369118X.2018.1428656

Edstrom, M. (2016). The trolls disappear in the light: Swedish experiences of mediated sexualised hate speech in the aftermath of Behring Breivik. *International Journal for Crime, Justice and Social Democracy*, *5*(2), 96–106. doi:10.5204/ijcjsd.v5i2.314

Elliott, M., Berenston-Shaw, J., Kuehn, K. M., Salter, L., & Brownlie, E. (2019). *Digital threats to democracy*. Retrieved from https://www.digitaldemocracy.nz/

Emanuelson, E. (2017). Fake left, fake right: Promoting an informed public in the era of alternative facts. *Administrative Law Review*, *70*(1), 209–232.

European Commission. (2018). *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

Farkas, J., & Schou, J. (2018). Fake news as a floating signifier: Hegemony, antagonism and the politics of falsehood. *Javnost—The Public*, *25*(3), 298–314. doi:10.1080/13183222.2018.1463047

Fletcher, R., Cornia, A., Graves, L., & Nielsen, R. K. (2018). Measuring the reach of "fake news" and online disinformation in Europe. *Reuters*. Retrieved from https://reutersinstitute.politics.ox.ac.uk/our-research/measuring-reach-fake-news-and-online-disinformation-europe

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, *10*(1), 33–50. doi:10.1386/jdmp.10.1.33_1

Foucault, M. (1972). *The archaeology of knowledge* (A. M. Sheridan-Smith, Trans.). New York, NY: Pantheon Books.

Foucault, M. (1991). Politics and the study of discourse. In C. Gordon, G. Burchell, & P. Miller (Eds.), *The Foucault effect: Studies in governmentality* (pp. 53–72). Chicago, IL: Wheatsheaf.

Fuchs, C. (2015). *Culture and economy in the age of social media*. London, UK: Routledge.

Fuchs, C., & Trottier, D. (2017). Internet surveillance after Snowden: A critical empirical study of computer experts' attitudes on commercial and state surveillance of the Internet and social media post-Edward Snowden. *Journal of Information, Communication and Ethics in Society*, *15*(4), 412–5444. doi:10.1108/JICES-01-2016-0004

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000233231

Galán-García, P., de la Puerta, J. G., Gómez, C. L., Santos, I., & Bringas, P. G. (2017). Supervised machine learning for the detection of troll profiles in Twitter social networking. *Logic Journal of the IGPL*, *24*(1), 42–53. doi:10.1007/978-3-319-01854-6_43

Gardiner, B., Mansfield, M., Anderson, I., Holder, J., Louter, D., & Ulmanu, M. (2016, April 12). The dark side of Guardian comments. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments

Gehl, R. (2018). Alternative social media: From critique to code. In J. Burgess, A. Marwick, & J. Poell (Eds.), *The SAGE handbook of social media* (pp. 330–350). London, UK: SAGE Publications.

Ghosh, D., & Scott, B. (2018). *#digitaldeceit: The technologies behind precision propaganda on the Internet.* Retrieved from https://www.newamerica.org/public-interest-technology/policy-papers/digitaldeceit/

Gillespie, T. (2018). Regulation of and by platforms. In J. Burgess, A. Marwick, & J. Poell (Eds.), *The SAGE handbook of social media* (pp. 254–278). London, UK: SAGE Publications.

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. doi:10.1080/1369118X.2019.1573914

Graeber, D. (2013). *The democracy project: A history, a crisis, a movement*. New York, NY: Random House.

Guess, A., Nyhan, B., & Reifler, J. (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign. *European Research Council*, *9,* 1–49.

Guo, L., Rohde, J. A., & Wu, H. D. (2020). Who is responsible for Twitter's echo chamber problem? Evidence from 2016 U.S. election networks. *Information, Communication & Society*, *23*(2), 234–251. doi:10.1080/1369118X.2018.1499793

Hille, S., & Bakker, P. (2014). Engaging the social news user. *Journalism Practice*, *8*(5), 563–572. doi:10.1080/17512786.2014.899758

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 40–60. doi:10.1111/jcc4.12001

Hintz, A., & Dencik, L. (2016). The politics of surveillance policy: UK regulatory dynamics after Snowden. *Internet Policy Review*, *5*(3). doi:10.14763/2016.3.424

Holcomb, J., Mitchell, A., & Purcell, K. (2015). Investigative journalists and digital security. *Pew Research Center*. Retrieved from http://www.journalism.org/2015/02/05/investigative-journalists-and-digital-security

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology*, *35*(2), 441–456.

Internet Governance Forum. (2015). *Recommendations on terms of service & human rights*. Retrieved from https://www.intgovforum.org/cms/documents/igf-meeting/igf-2016/830-dcpr-2015-output-document-1/file

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, *51*(8), 56–59.

Jakubowicz, A. (2017). Alt_Right White Lite: Trolling, hate speech and cyber racism on social media. *Cosmopolitan Civil Societies: An Interdisciplinary Journal*, *9*(3), 41–60. doi:10.5130/ccs.v9i3.5655

Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Bliuc, A.-M., Bahfen, N., . . . Connelly, K. (2017). *Cyber racism and community resilience: Strategies for combating online race hate.* London, UK: Palgrave Macmillan.

Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York: New York University Press.

Kamara, I. (2017). Co-regulation in EU personal data protection: The case of technical standards and the privacy by design standardisation "mandate." *European Journal of Law and Technology*, *8*(1), 1–24.

Kim, S., & Lee, J. (2012). E-participation, transparency, and trust in local government. *Public Administration Review*, *72*(6), 819–828. doi:10.1111/j.1540-6210.2012.02593.x

Klein, A. (2017). *Fanaticism, racism, and rage online: Corrupting the digital sphere*. London, UK: Palgrave Macmillan.

Klonick, K. (2015). Re-shaming the debate: Social norms, shame, and regulation in an Internet age. *Maryland Law Review, 75*(4), 1029–1045.

Lanois, P. (2016). Data protection in Singapore: What have we learned? *Journal of Internet Law*, *20*(2), 1–16.

Lewis, R. (2018). *Alternative influence: Broadcasting the reactionary right on YouTube*. Retrieved from https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf

Lu, J., & Yu, X. (2020). Does the Internet make us more intolerant? A contextual analysis in 33 countries. *Information, Communication & Society*, *23*(2), 252–266. doi:10.1080/1369118X.2018.1499794

Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society, 1*(2), 1–13. doi:10.1177/2053951714541861

Mackaskill, E., & Dance, G. (2013, November 1). NSA files: Decoded. *The Guardian*. Retrieved from https://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded#section/1

Marda, V., & Milan, S. (2018). *Wisdom of the crowd: Multistakeholder perspectives on the fake news debate*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3184458

Massanari, A. L. (2015). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346. doi:10.1177/1461444815608807

Massanari, A. L., & Chess, S. (2018). Attack of the 50-foot social justice warrior: The discursive construction of SJW memes as the monstrous feminine. *Feminist Media Studies*, *18*(4), 525–542. doi:10.1080/14680777.2018.1447333

Miller, C. (2016). *The rise of digital politics*. London, UK: Demos. Retreived from https://www.demos.co.uk/project/the-rise-of-digital-politics/

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society, 21*(7), 959–977. doi:10.1080/1369118X.2018.1444076

Mueller, R. S. (2019). *Report on the investigation into Russian interference in the 2016 presidential election*. Washington, DC: U.S. Department of Justice.

Nagle, A. (2017). *Kill all normies: The online culture wars from Tumblr and 4chan to the alt-right and Trump*. Winchester, UK: Zero Books.

Narayanan, A., & Reisman, D. (2017). The Princeton Web transparency and accountability project. In T. Cerquitelli, D. Quercia, & F. Pasquale (Eds.), *Transparent data mining for big and small data* (pp. 45–67). New York, NY: Springer.

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. New Media & Society, 20(10), 3720–3737. doi:10.1177/1461444818758715

Nguyen, A., & Vu, H. T. (2019). Testing popular news discourse on the "echo chamber" effect: Does political polarisation occur among those relying on social media as their primary politics news source? *First Monday, 24*(6). doi:10.5210/fm.v24i6.9632

Nielsen, M. M. (2017). eGovernance frameworks for successful citizen use of online services: A Danish-Japanese comparative analysis. *eJournal of eDemocracy and Open Government*, *9*(2), 68–109. doi:10.29379/jedem.v9i2.462

Noelle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication*, *24*(2), 43–51.

Oboler, A., & Connelly, K. (2018). Building SMARTER communities of resistance and solidarity. *Cosmopolitan Civil Societies: An Interdisciplinary Journal*, *10*(2), 99–118. doi:10.5130/ccs.v10i2.6035

Ombler, J., Russell, M., & Rivera-Munoz, G. (2016). Local councils and public consultation: Extending the reach of democracy. *Policy Quarterly*, *12*(4), 20–27. doi:10.26686/pq.v12i4.4623

Online Hate Prevention Institute. (2012). *Aboriginal memes and online hate*. Retrieved from https://ohpi.org.au/aboriginal-memes-and-online-hate/

Papacharissi, Z., & de Fatima Oliveira, M. (2012). Affective news and networked publics: The rhythms of news storytelling on #Egypt. *Journal of Communication, 62*(2), 266–282. doi:10.1111/j.1460-2466.2012.01630.x

Pariser, E. (2011). *The filter bubble: How the new personalized Web is changing what we read and how we think*. London, UK: Penguin.

Persily, N. (2017). The 2016 U.S. election: Can democracy survive the Internet? *Journal of Democracy*, *28*(2), 63–76.

Phelan, S. (2019). Neoliberalism, the far right, and the disparaging of "social justice warriors." *Communication, Culture and Critique*, *12*(4) 455–475. doi:10.1093/ccc/tcz040

Pohle, J., & Van Audenhove, L. (2017). Post-Snowden Internet policy: Between public outrage, resistance and policy change. *Media and Communication*, *5*(1), 1–6. doi:10.17645/mac.v5i1.932

Pöyhtäri, R. (2014). Limits of hate speech and freedom of speech on moderated news websites in Finland, Sweden, the Netherlands and the UK. *Annales–Series Historia et Sociologia Izhaja Štirikrat Letno*, *24*(3), 513–522.

Ray, R., Brown, M., Fraistat, N., & Summers, E. (2017). Ferguson and the death of Michael Brown on Twitter: #BlackLivesMatter, #TCOT, and the evolution of collective identities. *Ethnic and Racial Studies*, *40*(11), 1797–1813. doi:10.1080/01419870.2017.1335422

Risch, J., & Krestel, R. (2018). Delete or not delete? Semi-automatic comment moderation for the newsroom. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying* (pp. 166–176). Santa Fe, NM: Association for Computational Linguistics.

Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. In S. U. Noble & B. Tynes (Eds.), *Intersectional Internet: Race, sex, class and culture online* (pp. 147–160). New York, NY: Peter Lang.

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society*, *18*(2), 121–138. doi:10.1080/1369118X.2014.940365

Santana, A. D. (2014). Virtuous or vitriolic. *Journalism Practice*, *8*(1), 18–33. doi:10.1080/17512786.2013.813194

Shapiro, E. (2018). Point: Foundations of e-democracy. *Communications of the ACM, 61*(8), 31–34. doi:10.1145/3213766

Silverman, C. (2016, November 16). This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed News.* Retrieved from https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Sinclair, J. (2017). Political economy and discourse in Murdoch's flagship newspaper, The Australian. *The Political Economy of Communication, 4*(2), 3–17.

Sitrin, M., & Azzellini, D. (2014). *They can't represent us! Reinventing democracy from Greece to Occupy*. London, UK: Verso.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. doi:10.1016/j.jbusres.2019.07.039

Srnicek, N. (2017). *Platform capitalism*: Cambridge, UK: Polity Press.

Stohl, C., Stohl, M., & Ganesh, S. (2018). Digital media and human rights: Loomio, Statistics New Zealand, and gender identity. In A. Brysk & M. Stohl (Eds.), *Contracting human rights: Crisis, accountability, and opportunity* (pp. 232–249). Cheltenham, UK: Edward Elgar.

Stoycheff, E. (2016). Under surveillance: Examining Facebook's spiral of silence effects in the wake of NSA Internet monitoring. *Journalism & Mass Communication Quarterly*, *93*(2), 296–311.

Stoycheff, E., Liu, J., Xu, K., & Wibowo, K. (2019). Privacy and the panopticon: Online mass surveillance's deterrence and chilling effects. *New Media & Society*, *21*(3), 602–619.

Sweet, M., Pearson, L., & Dudgeon, P. (2013). @Indigenousx: A case study of community-led innovation in digital media. *Media International Australia*, *149*(1), 104–111. doi:10.1177/1329878x1314900112

Unsvåg, E. F., & Gambäck, B. (2018). The effects of user features on Twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online.* Brussels, Belgium: Association for Computational Linguistics.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Walton, S. C., & Rice, R. E. (2013). Mediated disclosure on Twitter: The roles of gender and identity in boundary impermeability, valence, disclosure, and stage. *Computers in Human Behavior, 29*(4), 1465–1474. doi:10.1016/j.chb.2013.01.033

Wizner, B. (2017). What changed after Snowden? A U.S. perspective. *International Journal of Communication*, *11*(5), 897–901.

Wu, W., Ma, L., & Yu, W. (2017). Government transparency and perceived social equity: Assessing the moderating effect of citizen trust in China. *Administration & Society, 49*(6), 882–906. doi:10.1177/0095399716685799

Ziegler, C. E. (2018). International dimensions of electoral processes: Russia, the USA, and the 2016 elections. *International Politics*, *55*(5), 557–574. doi:10.1057/s41311-017-0113-1

Zimmer, F., Scheibe, K., Stock, M., & Stock, W. G. (2019). Fake news in social media: Bad algorithms or biased users? *Journal of Information Science Theory and Practice*, *7*(2), 40–53. doi:10.1633/JISTaP.2019.7.2.4

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for the future at the new frontier of power*. London, UK: Profile Books.