DIGITAL THREATS TO DEMOCRACY

LITERATURE REVIEW PART 2: SOLUTIONS

Kathleen Kuehn, Leon Salter, Jess Berentson-Shaw & Marianne Elliott

The Workshop

The Law Foundation NEW ZEALAND

Luminate

DIGITAL THREATS TO DEMOCRACY

This report is part of the Digital Threats to Democracy research project.

To see the rest of the reports and the overall findings go to digitaldemocracy.nz

ISBN: 978-0-473-48026-4

CONTENTS

Following on from the review of the literature identifying the opportunities and threats that digital media pose to an inclusive and participatory democracy (Part One), we undertook a review to identify tested and workable solutions to realising the potential of digital media and/or overcoming current threats. Underlying this work is an understanding that ensuring an inclusive and fully participatory democracy is of such critical importance to our society that the current threats posed to our democracy from digital media requires us to understand what strategies, policies and behaviours we can use to achieve such an outcome and to act.

The review is presented in three parts: 1) the empirical evidence on workable solutions to threats to democracy from digital media, 2) a summary of recommendations found in the literature and 3) a brief discussion of some activities identified in New Zealand

A non-systematic narrative review was chosen with a view to summarising the evidence. Searches were limited to research published in the last eight years (most are within five). It was not an exhaustive review, but in general we found a dearth of empirically tested solutions. This dearth of tested interventions is not surprising given the slow response of government and other public institutions (from where such research would most logically be situated and or funded) to the threats from digital media.

PART 1. WORKABLE SOLUTIONS

In this part of the paper we present the interventions and solutions we identified in relation to each of the problems and threats that were identified in Part One of this literature review. We start with decreasing the power of private platforms.

DECREASING THE POWER OF PRIVATE PLATFORMS

Unfortunately little empirical evidence or evidence-based research has established what works for decreasing the power of private intermediaries. The dominant recommendations are generally calls for 1) more regulation (Betkier, 2018; Gillespie, 2018; Kamara, 2017; Marda & Milan, 2017); and 2) citizen-consumer activism, which ranges from pro-democracy to anti-capitalist initiatives (Srnicek, 2017; Zuboff, 2019).

The process for enacting these workable solutions also remain, at present, vague and non-specific. The suggestions are typically based on conclusions reached from multi-stakeholder interviews and focus groups as prescribed by legal scholars, academics, activists and industry leaders working in this space. Examples of recommendations to combat platform capitalism include:

Regulate platforms like other industries. Currently, regulatory debates largely centre around defining the structure, terms and conditions of what kind of industry private intermediaries represent. How platforms should be regulated or governed thus partly hinges on how these services are defined; for example, whether social media platforms are media companies, public spaces, utilities or some other service largely informs how they can ultimately be governed.

Introduce new modes of collective action (Zuboff, 2019). Under industrial capitalism we had collective bargaining, the strike – e.g., forms of collective action that were sanctioned by law and had support of society that allowed people to tame capitalism with legal protection. New forms of collective, collaborative action that connect users/consumers with the market and state to tame and outlaw surveillance capitalism must be invented. Zuboff's recommendations here map onto the work being done by advocacy groups like OHPI and autonomist Marxists (Hardt & Negri; Lazzarato; Virilio), who have been working to engage multiple vested interests lobbying for broader structural changes across the political economy *and* culture.

Creating a "sea change in public opinion" (Zuboff, 2019) that will no longer tolerate "surveillance capitalism" as the dominant economic form that trades in human futures, or one where government dips into servers held by private intermediaries for surveillance purposes. Zuboff believes a sleeping democracy has allowed these companies to create asymmetries of knowledge and power antithetical to democracy and that a shift in public attitudes is needed to persuade these companies to change.

Design New Competitive Solutions (Jackson & Kuehn, 2016; Zuboff, 2019). Dissatisfaction with large intermediaries' power presents new business opportunities that can forge an alternative digital future that in turn, facilitates a more democratic internet. Zuboff believes growing public dissatisfaction provides a space for disruptive technology to emerge; that is, platforms offering a different set of techno-social affordances have an untapped market waiting to be exploited. Platform cooperatives like Loomio (Jackson & Kuehn 2016; Ombler et al., 2016; Sandoval, 2016; Scholz, 2016; Stohl et al., 2018), subscription-based models and pro-privacy and non-commercial alternatives (Beattie, forthcoming) are already in use but have not reached the critical mass needed to become a competitive solution just yet.

Improve Content Moderation (see ' Recommendations' section below).

FOREIGN GOVERNMENT INTERFERENCE IN DEMOCRATIC PROCESSES

The research in this area is also largely normative, but seems to generally prescribe the design of new anti-cybersecurity infrastructure that will reduce threats to elections and other political processes (Hicks, 2018; Shoorbajee, 2018). Drawing upon "big datasets" to review and assess electoral policies, processes and outcomes – including the potential interference from foreign governments – is another common recommendation (The Election Administration and Voting Survey). Generally, most call for multi-stakeholder input on solution design, but proposals largely seem to emphasis technical over human solutions.

SURVEILLANCE & DATA PROTECTION

The opacity of mass and commercial surveillance presents challenges to empirical research looking to find and test workable solutions to the threat it poses to democracy. Some regulatory measures, like the Singaporean Data Protection Act 2012, work to regulate companies' information management practices and have been proven effective in bringing formal charges to data mismanagement and abuse. Singapore's regulatory measure created a Data Protection Authority (DPA) that it invested with the power to: conduct investigations on the data collection policies and practices of organisations; order the destruction of data; and impose fines for data mismanagement and abuse (Lanois, 2016). (The act extends to companies that are not based in Singapore, as well). By 2016, the DPA had received 667 complaints, took action against 11 organizations for data breaches and lodged four financial penalties (Lanois, 2016). One particular prosecution against smartphone manufacturer Xiaomi resulted in the company making changes to the terms of its cloud-messaging service agreement and data storage practices.

A majority of research, however, calls for regulatory changes to data privacy policies (e.g., Fuchs & Trottier, 2017; Internet Governance Forum, 2015; Flew, 2019). However, there is little evidence to suggest that these changes will reduce surveillance/data collection so much as regulate how that data is stored, accessed and used by data collectors and other third parties. Instead, privacy advocates typically encourage internet users concerned with reducing online surveillance or mitigating the anxieties and concerns it engenders by employing technical solutions like ad-blockers and ad-tracking browser extensions, private browser options (e.g. Tor), open source platforms and cooperative platform models all present alternatives to subvert surveillance mechanisms (Narayanan & Reisman, 2017). "Evidence" supporting the efficacy of these tools and alternatives, however, is typically anecdotal or prescriptive in nature (as opposed to empirical), and is based on the premise that these alternatives increase public awareness of tracking, bring it into public debate and in turn, decrease some of the more invasive surveillance practices.

FAKE NEWS

Few studies on fake news have yet to offer concrete empirical solutions for combatting this phenomenon. The dominant mode of published research is largely normative in approach, making prescribed recommendations for potential solutions to combatting fake news, yet none of which have been tried or tested. Methodologically, these recommendations are typically the result of multi-stakeholder interviews with experts across the public and private sectors (e.g., Marda & Milan, 2018) or joint declarations from international organisations (e.g., OSCE, 2017). Examples of proposed – but untested – recommendations to combat fake news starting at the structural level and moving towards individual level include:

- Supporting a vibrant and diverse media sphere that balances strong, independent and adequately resourced public service media with a non-concentrated commercial media sector (OSCE, 2017);
- Developing and circulating persuasive counter-narratives [with emotional, versus rational, appeal] (Dubow, 2017; OSCE, 2017)
- A multi-stakeholder approach to content moderation that combines human and technical intervention (Klonick, 2018; Madra & Milan, 2018)
- Education, particularly around critical thinking (Dubow, 2017)

FILTER BUBBLES/ ECHO CHAMBERS

PLATFORM DESIGN & AFFORDANCES

Technically speaking, affordances refer to the perceived range of possible actions related to the features of any given platform (Bucher & Helmond, 2018; Hutchby, 2001). In nontechnical language design affordances are invitations, prompts and clues on how to use a thing. In digital media they ascribe meaning to how to use the tool, for example Facebook has a "friend" button directing the user towards ways of interacting based on mutual agreement, also a "share" button, while Twitter has a "follow" button, open to all people using the platform, directing or suggesting different ways of interacting. Other examples includes the use of anonymous accounts, retweet buttons, mentions or share data feedback. The design of these affordances has an impact on inclusion and participation, as well as the types of interactions people experience and information they are exposed to.

There is some suggestion that design affordances can reduce the effects of filter bubbles by engaging internet users in more ideologically diverse communities. While designing features to encourage or delimit certain behaviours or activities in an online space does not guarantee they'll be taken up that way by users (and in fact, users often take up features in ways unintended or unanticipated by designers and platform owners), platform features have also been shown to effectively direct or delimit action in certain ways (Bucher & Helmond, 2018).

Non-commercial platforms like Loomio, for example, afford different modes of interaction based on the features (e.g., tools, interface) and environment (e.g., deliberative; asynchronous) it makes available outside a commercial space. Unlike privately owned social media platforms, user identities are less curated towards a consumerist framework,

rather the platform affords an environment of less performative, and thus more considered, dialogue, discussion and debate. The platform's design at least partly contributes to its effective use by government agencies, local councils and advocacy groups to consult and deliberate with citizens on a range of contentious initiatives and strategies.

The relationship between design and civility is demonstrated in a study that analysed The Wellington City Council's use of Loomio to debate proposed changes to alcohol licensing rules, which the Council ruled as a success in more creating in-depth discussion with less polarization than offline debates. Getting constituents to meaningfully engage outside echo chambers may be partly tied to Loomio's time-based affordances. The limited time given to public face-to-face meetings often creates pressure for speakers to get their points across forcefully and enhance selective bias by creating discursive alliances between constituents who seemingly share similar opinions. Loomio, however, "opens up space where all views can be considered and everyone can still be heard" without time-based anxieties and restrictions (Ombler, Russell, & Rivera-Munoz, 2016, p. 23). Loomio's platform affordances also move deliberation beyond *debate* to collective *agreement* (i.e., rational consensus) with the goal of making "the fewest people unhappy" (Rushkoff, 2014, n.p.).

These kind of well-designed, collectively-owned, online deliberative fora not only offer a safe place for different views to be expressed and heard (effectively reducing the propensity to engage with similar-minded people encouraged by automated filter bubbles), but offers a safe space for marginalized groups, as well. Loomio was effectively used by Stats NZ to debate the inclusion of "gender diverse" as a third category.[1] Despite criticisms surrounding the process, an analysis of feedback left on the Stats NZ website reported the process as largely positive, which contributors "recognized to be an inclusive discussion" (Stohl, Stohl & Ganesh, 2018, p. 246). Further, it gave "visibility to an issue that needed to be made more public" and "brought people together into the conversation who had previously been marginalized" (p. 246). As detailed in Part 1, hate speech instigated by members of a vocal majority can curtail the voicing of opinions by minority groups, pushing them further into the margins, thereby narrowing the public sphere.

1. The decision which was later rescinded, however, "for purely statistical reasons" (Stats NZ, 2018).

HATE SPEECH/ TROLLING

IDENTITY VERIFICATION SYSTEMS

While hate speech and trolling actually constitute a small minority of online comments (Papacharissi, 2004; Rowe, 2015), attempts to combat intolerance, incivility and other forms of disruptive speech have a long history in internet research. To date, however, there little consensus on efficacious strategies.

The role of anonymity online has been perhaps one of the most commonly sites of analysis for mitigating hateful, uncivil and impolite behaviour online. While early scholars advocated on behalf of anonymity's liberating affordances, namely around identity play and self-expression (e.g., Baym, 2012), the dominant view in empirical research positions anonymity as directly related to increased incivility online. Trolling, for example, generally hinges on abuse by anonymous individuals with "fake" accounts.

Sites that force pre-registration have been shown to solicit qualitatively better, but quantitatively fewer, user comments because of the extra effort required for engaging in discussion; forced registration can also facilitate 'known' user identities, which can serve to hold users to account (Bakker, 2010; Rowe, 2015; Santana, 2014). Quantitatively, when comparing user comments posted to the same stories on a news outlet's website and its corresponding Facebook page, Internet users are far more likely to comment on the website version of news stories than similar stories featured on the organisation's Facebook page (in both studies, website comments enabled anonymous posting, whereas Facebook does not).

Empirical research has also found that abusive comments are minimised when anonymous commenting is prohibited (Santana, 2014) or when identity verification is required through systems like Facebook comments (Rowe, 2015). One study, however, found that despite the ability to remain anonymous, comments on news websites were of comparatively higher quality to those on Facebook and sparked more vibrant debate, whereas Facebook comments contributed little engagement; the authors concluded that "Facebook will provide few comments, will kill the trolls, but will not result in making the conversation more interesting" (Hille & Bakker, 2014, p. 572).

Experimental research similarly shows that identification-based systems like Facebook comments correlate to more civil forms of online discourse than on platforms in which comments are posted anonymously on the same news stories (Rowe, 2015). Rowe compared news comments posted to the same stories featured on *Washington Post's* website to comments posted to the newspaper's Facebook page and found the rate of uncivil behaviour to be significantly higher on the former (where comments were anonymous) than on the its Facebook page (where comments are not anonymous). The study also found incivility to be overall more personally insulting and directed at specific users participating in political discussion on the newspaper's website. On Facebook, however, incivility and impoliteness were "aimed at individuals not involved in the discussion, or used as a way to articulate an argument, rather than offend others" (p.132). Rowe concluded that visibility to one's wider social network afforded by Facebook ostensibly holds users to account for their communicative actions in ways that website's anonymity does not.

Facebook's affordance of social surveillance thus functions as a sanctioning mechanism that enables more civil and tolerant political discussions. The conclusion, then, is that anonymity encourages more abusive behaviour while visibility or forced identification can enable more civilised forms of discourse.

PLATFORM DESIGN & AFFORDANCES

De-anonymising internet communications speaks to the role that platform design can play in mitigating incivility and hate speech online. A system or platform's communicative and social affordances at least partly explain how and why some cultures of participation evolve the way they do across different social media platforms. While user characteristics, user goals/objectives and cultural context also significantly influence how platforms are engaged, there is some evidence that suggests well-designed, digital fora can encourage stronger deliberation practices, mutual respect and accountability through their discursive promotion of implicit and explicit rules of conduct (Kavada, 2012). Research suggests that intentionally building more participatory forms of engagement into platforms might reduce filter bubbles, echo chambers *and* incivility (particularly on mobile devices), while increasing communication and deliberative processes (Groshek & Cutino, 2016; Halpern & Gibbs, 2012; Hmielowski, Hutchens, & Cicchirillo, 2014).

Therefore, the act of *consciously* designing social platforms to engender pro-social forms of engagement can have a demonstrated impact on civility. However, so also is the act of *consciously selecting and choosing* a particular platform or set of tools over others; for e-government initiatives in particular (see section on distrust/dissatisfaction with democracy), carefully selecting what (and how) social media platforms are engaged in relation to the overall objective can assist organisations, governments and other actors in facilitating more deliberative, civil forms of online discourse, as well. Writing code (e.g., features; interface design; verification systems) that enables pro-social forms of engagement and intentionally disables or limits anti-social behaviours and actions is thus one workable solution that engineers can take up as they develop and innovate new communicative platforms.

CIVIL SOCIETY & ADVOCACY ORGANISATIONS: BUILDING RESILIENCE THROUGH SUPPORT NETWORKS

Developing fast and effective reporting mechanisms and support networks are another set of workable solutions thought to combat hate speech online. Advocacy and civil society organisations like All Together Now, the Federation of Ethnic Communities' Councils of Australia (FECCA), and Australia's Online Hate Prevention Institute (OHPI) have demonstrated some success with building online reporting tools that rely on crowdsourcing to identify – in order to remove - racist hate speech online (Bodkin-Andrews, Newey, O'Rourke, and Craven, 2013, Jakubowicz et al., 2017, Sweet et al., 2013; Oboler & Connelly, 2018). For example, the OHPI's website FightAgainstHate.com offers a reporting tool that gathers and compiles evidence on the extent of the issue while also highlighting moderation gaps amongst larger online platforms.

Elsewhere, case studies of online communities like @IndigenousX (a "grassroots Twitter community made up of Indigenous Australian guest tweeters and followers") have also shown that a networked approach can effectively combat the effects of hate speech; in this case, @IndigenousX exemplifies a growing base of "resilient communities" that act as support networks to individuals who come under racist attack (Jakubowicz et al., 2017, p. 236). These groups can effectively use Twitter as a form of "participatory journalism" to build counter-narratives that counteract racism directed at indigenous Australians or others (Sweet et al, 2013).

Other research has shown the role that Twitter hashtags play in providing a space for the construction of resistant collective identities in cases like the #BlackLivesMatter movement or #MeToo (Ray, et al., 2017). The extent to which victims of online abuse report higher levels of 'resilience to racism' through these tactics and support networks, however, has not been empirically tested to our knowledge.

Coordinating diverse stakeholders to apply pressure to private intermediaries, particularly in 'long-haul' campaigns, has also been effective in having hateful content removed from social media, even when content does not technically violate a platform's terms of service (ToS). Facebook eventually conceded to the removal of content and imagery expressing hatred against Aboriginal Australians after pressure from the OPHI, not because the content violated Facebook's ToS but because a broad and diverse range of actors found the content insulting and complained, which included civil society, advocacy groups, regulators and individual users. Sustained pressure from diverse stakeholders tends to garner mainstream media attention that threatens the brand reputation of platforms can have a demonstrated effect on the forced removal and moderation of hateful online content ("Aboriginal Memes & Online Hate", 2012; Gargliardone et al., 2015). Speed of removal is considered essential to diffusing the power of hate speech and trolling, as the longer hateful content remains online, the more damage it inflicts on victims while empowering the perpetrators (OPHI). Early content removal thus effectively limits the scale of exposure.

Pressure from researchers and advocacy groups alike have also encouraged some platforms to design more pro-social tools (i.e., affordances) into their systems. Twitter has responded to public pressure by unrolling a number of features that afford users the opportunity and ease to contribute to content moderation. These features include the ability to hide content users do not wish to see; the ability to flag harmful or inappropriate content ("Report Abuse" buttons); verified accounts aimed at building trustworthiness; a "quality" filter that uses an algorithm to hide spam-like tweets; a "notifications" filter for disabling notifications of mentions or replies from people they don't follow and so on (Klonick, 2016). While these features afford users the opportunity to intentionally or semi-automatically moderate harmful and uncivil content, they also threaten to oversanitise online spaces, intensify filter bubbles/echo chambers and to disconnect users from wider network affordances (e.g., diversity, new followers, etc).

CONTENT MODERATION POLICIES & PRACTICES

Calls for new regulatory policies around content moderation at large intermediaries are gaining traction as a necessary means of combatting both incivility and misinformation online. This remains the case even in light of the growing consensus that content moderation remains an opaque and difficult practice, and on its own is not a fix-all solution.

On the one hand, the internet's global reach presents the basic challenge of moderating what constitutes "appropriate" content across vastly different cultures marked by their own distinct values and standards, which create variance across how even 'objective' content rules and policies are interpreted and instituted (Roberts, 2016). Contextual issues also make quantifying inappropriate content difficult; e.g., resolving the debate over why one piece of content is acceptable but a slight variation breaks policy in another is the "holy grail" of moderation (Diakopoulos & Naaman; Klonick, 2018; Pöyhtäri, 2014).

On the other hand, content moderation policies at the largest intermediaries remain largely opaque, although what scholars have uncovered is that most policies attempt to balance stakeholder expectations (including users, consumers, advertisers, shareholders, the general public), commercial business goals, and jurisdictional norms and legal demands (which are generally governed by liberal-democratic notions of "free speech") (Gillespie, 2018; Klonick, 2018; Roberts, 2016). As Roberts (p. 9) summarises, "Internal policies regarding permissible content therefore serve these purposes, first and foremost, rather than responding to, say, social justice or advocacy-related goals."[2] Typically, content moderations processes combine company resources (in the form of internal and external content moderators; or technical solutions like automated filters) with user labour (e.g., via flagging or other online reporting mechanisms). These processes can be reactive or proactive (sought after); manual or automated.[3] Some platform operators source experts or trained specialists in suicide, human trafficking, child exploitation, domestic violence, terrorism, while others conduct semi-regular audit reviews of moderator decisions to ensure consistency and adherence to policy guidelines.

The most common 'workable solution' presented as it relates to content moderation are processes that combine technical and social (human) responses. The manual policing and removal of online hate speech has a long history on news websites, where community managers, employees and journalists have taken up for years as part of their public interest mandate (Pöyhtäri, 2014). Flagging and removal by users has also been a demonstrably effective, albeit labour-intensive, way of having nefarious and offensive speech removed. On the technical side, however, advances in semi- or fully automated systems, including deep learning, show increased promise in identifying inappropriate content and drastically reducing the number of messages human moderators then need to review (Binn et al, 2017; Delort, Arunasalam & Paris, 2011). Tested technical solutions include Reddit's 2015 ban on two subreddits rife with hate speech (r/fatpeoplehate and r/CoonTown) (Chandrasekharan et al.'s (2017). The study found Reddit's use of an automatic keyword identification tool effectively encouraged hate-speech accounts to discontinue their use of the site, while remaining accounts reduced their hate speech use by 80 percent. Elsewhere, Galán-García et al. (2014) tested their machine-learning algorithms have been used to track cyberbullies trolling their peers at a Spanish school, effectively narrowing the perpetrators down to three students who then confessed to the bullying. However, these types of tracking techniques raise a number of concerns around algorithmic sorting and institutional surveillance, particularly in educational settings.

Other ongoing research aims to advance a more holistic approach that semi-automates content moderation via more transparent classification systems that try to account for context while providing moderators and users a reason for their classification results (e.g., providing users with an explanation for content deletion) (Risch, J., & Krestel, 2018). Researchers have also found text-based mining alone to be insufficient, and have thus turned towards non-text features like user characteristics as potential datasets for detecting incivility online. In one study, combining certain user features with textual features slightly improved the performance of automated classification results in hate speech detection models (Unsvåg, 2018; Unsvåg & Gambäck, (2018). This tactic again, however, functions on the submission of users to more surveillance.

Empirical researchers are increasingly willing to admit, however, that neither automated nor manual classifications systems can ever be "neutral" or free from human bias: "There can be no formula determining the extent to which different viewpoints need to be

2. Klonick's historical and qualitative research with insiders tied to Facebook, YouTube and Twitter found finds that platforms developed their moderation systems in accordance with (1) an underlying belief in American constitutional free speech norms; (2) a sense of corporate responsibility (tied mainly to enforcing democratic norms); and (3) the necessity of meeting users' norms for economic viability (which dominates any sense of corporate responsibility). This is echoed by scholars like Sarah T Roberts, who found content moderation processes at large social media firms are governed by policies that similarly try to balance (1) attracting user-participants and advertisers; 2) responding to jurisdictional norms and legal demands, and 3) remaining profitable and appealing to shareholders.

3. The dominant approach to moderation is based on a human-generated rules-based approach (e.g., white lists; black lists) that are manually processed and resource-consuming; they are not only prone to human bias but often produce erroneous results (Gillespie, 2018; Delort, Arunasalam & Paris, 2011). These automated process are also "difficult to maintain as language, norms, and gaming strategies change" (Binns, Veale, Van Kleek, Shadbolt).

reflected in order for a classifier to be deemed fair" (Binns et al, 2017, p. 411; see also Gillespie, 2016). Human and/or automated content moderation is unlikely to achieve "civil discourse," a "sanitized" internet or other speech and engagement goals through moderation alone. Therefore, the combination of automated classification and deletion systems and human efforts remains the most effective content moderation strategy currently on offer. Even still, the complex technical and cultural issues that undergird effective moderation have not stopped claims that "more moderation: is the 'answer' to combatting the internet's misinformation campaigns and speech problems. In the few places where they exist (German Network Enforcement Act, 2017), government regulations on private intermediaries' moderation practices have not been empirically tested for their efficacy or effectiveness.

DISTRUST/ DISSATISFACTION WITH DEMOCRACY

Increasing trust in government institutions (and democracy broadly) is covered by a range of literature on direct and participatory democratic engagement/processes, e-government, and open government.

The creation, selection and use of online platforms that afford citizen participation and deliberation can also enhance government trust (OECD, 2017; Valtysson, 2013). On the one hand, governments need to offer opportunities for citizens to engage in all levels of policy development, including their "design, implementation, monitoring and evaluation" (OECD, 2017, p. 118); such initiatives have the potential to engage and empower all citizens, while "forging a new relationship between cities and local residents, and parliamentarians and citizens" (Simon et al, 2017, p. 11). However, the systems and platforms used in these deliberative processes must also enable these practices to emerge. As mentioned, the techno-social affordances inherent to different online platforms affect and shape the nature of engagement, deliberation and discussions (Bucher & Helmond, 2018; Valtysson, 2013). For example, a recent case study on the use of social media to 'crowdsource' the re-writing of the Icelandic Constitution provides empirical support for how social media generates different user responses and modes of engagement. An analysis of this process found that the interactive and participatory affordances of varying social media enabled Council members to provide practical information with constituents, disseminate information to mainstream media and to maintain a positive rapport with citizens via enthusiastic responses to their comments, discussion and suggestions. Yet the study also found that social media – and the government's use for deliberative democracy – fell short as a space to effectively achieve rational consensus. Social media might *afford* state actors to constitute and engage networked publics, but the practice of extending deliberation to activate the public as a decision-making body requires using government to encourage platform use in this way.[4]

A meta-analysis of empirical research found that contributions to democratic modes of participation differ according to type of democratic innovation: e.g., deliberative forums and surveys better promote exchange of arguments, "whereas referendums and participatory policy making projects are better at giving citizens influence on policy making and involving more people" (Michels, 2011, p. 275). Indeed, international research has found that engaging citizens in deliberative processes often results in profound changes in deliberating citizens' "frequently in the direction of more common good-

4. The author concedes social media can engender strong, effective and civil political *discussion*, but operationalises 'deliberation' by the Habermasian standard of achieving rational consensus – the point at which social media falls short.

oriented policies" (Bächtiger et al., 2011, p. 5). This finding seems to stand even in the context of highly polarised issues, although the influence of deliberation on preference transformation towards less simplistic measures is also thought to be moderated by a period of internal reflection (Bächtiger et al., 2011). This period of "deliberation within" may in fact be more important to preference or opinion change than discussion itself (Goodwin, 2003), suggesting that time to adequately reflect upon information provided before engaging in deliberative discussion is more important conversation itself. Empirical research supports that preceding online modes of deliberation with a period of internal reflection is effective, yet quite the challenge in a communicative environment premised on immediacy, constant updates and a dynamic information cycle. It's worth noting as well that 'deliberation' in these research contexts are typically operationalised as engaged, civil discussion, which precludes Habermas' requisite of achieving rational consensus.

Transparency, access and design are essential to generating positive results from digital government initiatives, and when done effectively have been shown to increase positive feelings and citizen trust in local government (Leininger, 2015; Kern, 2017; Swaner, 2017). The shift towards "e-government," whereby information and communication technologies (ICTs) are used to improve public sector activities, aims to improve state services and engage more citizens in democratic processes (eGovernment for Development, 2008; OECD, 2017). Similarly, 'open government' initiatives are designed to increase transparency, inclusiveness and thus institutional trust, but have largely advocated for "open data" initiatives that make government data freely available for public use – a insufficient solution on its own (Lourenço, 2015). Instead, adequate design, resourcing and accessibility are central to successful e-government and open government initiatives. Some evidence supports the finding that governments that have created usable, intelligible websites, and offer non-exclusionary solutions for those lacking computer and internet access or basic digital literacy skills, have been most successful in their e-government initiatives and constituent satisfaction (UN, 2018).

Finally, civics education, or educating children in schools on "good citizenship" has been positively associated with increased political engagement, particularly for socially and economically marginalised groups (Neundorf, Niemi & Smets, 2016; Van de Werfhorst, 2017).

PART 2. COMMON RECOMMENDATIONS ACROSS THE LITERATURE:

This research confirms that there is, at present, a troubling dearth of scientific, empirical, evidence-based scholarship that tests or aims to validate "workable solutions" to the seven key threats to democracy we've identified in this project. A significant majority of the academic research relating to these seven problems is *normative* in approach; meaning, it presents theoretically sound arguments about the way things "ought to be" if democracy is to be "reclaimed" from incivility and a rogue form of capitalism in the digital age.

While there is certainly a place in both academia and policy for normative (and critical) research, there is also a need to begin testing many of the propositions and recommendations being made and instituted by industry, governments, NGOs and civil society to not only measure and extend what's working to other places or contexts, but so that future normative prescriptions are informed by evidence beyond the anecdotal (or budgeting restrictions), as well. We need to start creating testable knowledge beyond good ideas and theory-informed 'recommendations' (although again, there is a place for that).

What follows is a summary of prescribed recommendations that appeared most frequently across the literature as advocated by scholars, legal experts, advocacy groups and other civil society organisations. It is not an exhaustive list. It is also worth noting that many of these recommendations emerge from – and also recommend – a multi-stakeholder approach to internet governance that represents and balances the interests of industry, state, and civil society. In such cases we must consider carefully how people from civil society are included in a way that take account of the power imbalances between this group and the other two.

1. POLICY / LEGAL SOLUTIONS

Proposals here range from calls to review and overhaul existing legislation; create new legislation; institute new oversight bodies or inter-government agencies (e.g., Establishing an independent Commissioner and/or oversight board with the power to impose fines on private intermediaries; remove social media from ISPs; regulate/vet algorithms to minimise injury to public interest); or to improve regulations on content moderation.

In the US context, many scholars are calling for regulators to revisit and consider overhauling the governing legislation around moderation issues (Gillespie, 2018; Klonick, 2018; Laidlaw, 2015). This legal framework is embedded with distinctly American normative values around free speech and free enterprise that have significant ramifications for the rest of the world (particularly places that do not share these same cultural values). Currently, the governing legislation for social media content is Sec 230 of the 1996 Communications Decency Act, which absolves all internet companies and intermediaries – from ISPs to platforms to content makers – from liability for any content they host while allowing them to also delete any content they wish without consequence. By this law, platforms are not legally obliged to remove or filter content that does not violate the company's terms of service, even in cases where legal courts find such content defamatory (e.g. Hassell vs Bird, 2018). Conversely the rule simultaneously allows platforms to remove or filter content that constitutes 'free speech' or some other constitutional protection without consequence. Revising this regulation, however, first requires determining whether platforms constitute media companies (broadcasters or editors?); state actors; a "public" or public town; a (public) utility or some hybrid formation for which current speech regulations do not yet exist or apply. How private intermediaries are defined by law thus inform their regulation, yet these are questions that have only begun being asked let alone resolved.[5]

2. MORE CORPORATE TRANSPARENCY

Currently the lack of transparency around moderation practices presents challenges to accountability, governance, and the ability to apply public and legal pressure. Effective examination of the "moderation apparatus" includes access to private intermediaries' a) content moderation policies; b) the sociotechnical mechanisms used in their enforcement (e.g., system design, labour, organisational culture); c) business expectations the apparatus must serve; d) the justifications articulated to support these criteria (Gillespie, p. 12-13). Expanding empirical research to improve moderation processes requires private intermediaries make these processes and practices accessible to researchers.[6] Access to formal documents or policies related to moderation processes are also generally unavailable, and in cases where semi- or fully-automated techniques contribute to content moderation, little insight is available about how these processes work, their efficacy or the algorithms behind them.

5. For example, Klonick (2018) advises against treating platforms as state actors bound to First Amendment (free speech) law given the difficulties of determining "when a private party's behaviour constitutes state action," e.g., in what situations or context private property functions as a public space (p. 1659). Regulating as media companies is also a tenuous proposition as in the US, the regulation of broadcasters is based on redressing matters arising from spectrum scarcity and "invasive" nature that don't apply in the same way to the digital context (e.g., the lack of spectrum space justifies need to ensure media meet public interest needs, although it's certainly possibly that the right to platform access might challenge future scarcity claims). Treating platforms as 'forums' or town squares would give them their own First Amendment speech rights, which invests platforms with what some believe is too much independent regulatory power.

6. Research has generally found that the state of platform content moderation by private intermediaries is a closed, fractured and private process inaccessible to users, the public and even researchers; these conditions at least partly account for the lack of empirically-based research on content moderation practices. Despite calls for more transparency, content moderation remains a closed process, if not protected as an industrial trade secret. Human content moderators are themselves sequestered into silence through non-disclosure agreements (NDAs), fragmented and siloed as their own internal department (separate from engineers or company workforce) if not outsourced to private "boutique" firms, overseas micro-labor sites or call centres in the form of low-waged, low-status labor. In the latter case, moderation jobs are often cast as "customer-service" managed and staffed by women, which stands apart from the higher-status, higher-paid, more powerful sectors of engineering and finance, which are overwhelmingly male.

3. BETTER DESIGN

The consideration of platform affordances shows how design can influence the way individuals, organizations and institutions make decisions around platform uses/objectives. Pro-social and democratic values must be encoded into the infrastructure of the internet, including algorithms (Annanny, 2016; O'Neil, 2016). Assessing what normative values are being built into an ever-evolving internet infrastructure, whose values they represent and what forms of social life they afford are primary research questions guiding this approach. At present, the normative values embedded into these global private intermediaries – e.g., openness, connectedness, free speech, etc. – are not culture-neutral norms. While this creates a range of problems for an internet governed by global private intermediaries (including wholesale censorship regimes like China and North Korea's banning of Facebook and Twitter), it is also the first step towards designing more deliberative spaces, pro-social tools and online environments.

4. IMPROVE CONTENT MODERATION

Calls range from the standardisation of industry-wide "best practices" (e.g., Harvard's Berkman Center for Internet and Society, 2019) to more transparency and researcher access (Gillespie, 2018; Roberts, 2016). Suggestions to the former recommendation would require: a) Corporate transparency, consistency, clarity, and clear mechanisms for customer recourse around moderation complaints; b) Corporate grievance mechanisms that are transparent, accessible and in accordance with international human rights law; c) A multi-stakeholder governance approach that demands platforms engage in public dialogue with relevant organisations (Anti-Defamation Leagues, Digital Rights Foundation, National Network to End Domestic Violence, etc). (Gillespie, 2018; Harvard's Berkman Center for Internet and Society, 2019). Suggestions also include making content moderation an organisational priority rather than department silo, and one that brings together legal, customer service, security, privacy, safety, marketing, branding, and personnel to create a unified approach to resolving such a complex issue (Gillespie, 2018; Klonick, 2018; Roberts, 2016).

PART 3. ACTIVITIES IN NEW ZEALAND

Given that in our literature review we identified very little empirically tested evidence of any quality, and a plethora of expert opinion, we also looked to identify activities relevant to digital democracy in New Zealand. We did not do so systematically, and we have little or no evidence as to their impact on optimising opportunities, reducing threats, or improving democratic participation. This section therefore is purely a high level scan of local activities.

1. Two 2016 reports (Open Data NZ, 2016a, 2016b), describe public engagement measures that have been undertaken in New Zealand. For example, The Open Data Charter Public Consultation "ran throughout August and September 2016...designed to give interested and potentially affected parties with the opportunity to provide their thoughts about whether New Zealand should adopt the Open Data Charter and the impacts if it is was to be adopted" (Open Data NZ, 2016b, p. n.p.). Further, a nationwide survey was conducted, "asking people what the top 10 datasets they would like to see released were", to which they received "over 400 responses" (n.p.).

2. We have previously discussed two Loomio experiments conducted by Wellington City Council to consult on an alcohol management strategy, and Statistics New Zealand on gender categorization.

3. Work is currently being undertaken by Government Information Services (2018) in the area of e-government inclusivity. For example a Digital Inclusion Ministerial Advisory Group had been established (InternetNZ 2018). and the Government Online Engagement Service (GOES) was established in 2011, which has been piloting a survey tool asking people "how they would like to have their say with government" (Government Information Services, 2018, p. 11). Results suggest a thirst in New Zealand for initiatives in digital participative democracy.

4. The New Zealand government developed a cyber security strategy in 2012 (updated in 2016), and established the National Cyber Policy Office in the same year, and Netsafe in 2016, under the Harmful Digital Communications Act. (InternetNZ, 2017a).

5. A review of the Privacy Act by the Privacy Commissioner, which includes a focus on data protection for consumers (Privacy Commissioner, 2016);

6. The monitoring of the Harmful Digital Communications Act by the Human Rights Commission, in order to ensure we "strike the right balance between freedom of expression and our need as a community to challenge our bullying culture and protect people who are under attack" (Human Rights Commission, 2015).

7. A submission to the United Nations committee that oversees the Convention on the Elimination of Racial Discrimination, in order to improve the Human Rights Act to better account for online hate speech (Human Rights Commission, 2017).

8. The Local Government Excellence Programme (LGNZ, 2016) is a program which councils can choose to sign up to voluntarily, and provides them with an independent assessment "across four priority areas that the general public and business communities have told us are important to them" (p. 3). Councils are then ranked on their performance. This programme was specifically designed to counter declining trust in local councils. While the programme was only established in 2016, the 2017 Local Government Survey (LGNZ, 2017) showed a slight improvement in overall satisfaction.

9. Design+Democracy and RockEnrol, academic and civil society organisations, have undertaken work in the area of youth engagement. Design+Democracy's (2017) On the Fence describes itself as a "gameful questionnaire" and a "fun educational tool that helps young undecided and first-time voters engage directly with issues by matching their personal values with political candidates and parties". In the six week period before the 2017 election day, the site received "170,000 unique visitors", who, on average, "achieved a 92% completion rate". "A Horizon Research survey showed that the site encouraged 30,000 non-voting 18–34 year olds to vote, representing 7% of the total eligible youth population".

CONCLUSIONS

While our literature review was not exhaustive, the general finding was that there is a sheer dearth of empirical evidence in this area. This lack of evidence of any quality leads us to conclude that there is a critical need for people in government, civil society, NGOS and enterprise to invest in researchers and projects who will do pre- and post-testing of the solutions that stakeholders are recommending, including any recommendations the government decides to take up. It is critical that people in the New Zealand Government see whether or not what's being done / put in place is working. New Zealand would break significant ground in that regard.

Undertaking research in the following areas is critical:

- Understanding and agreeing what social media platforms constitute. Are they media companies (broadcasters or editors?), telecommunication companies, state actors, a "public" or public square, a utility, or some hybrid of the above? And given this, what policy and legal solutions work best to minimise public harm?
- Exploring the impact of public investment in new competitive solutions to forge an alternative digital future and facilitate a more democratic internet.
- Determining whether the creation and support of new modes of collective action can achieve broader structural changes across the political economy and culture.
- Understanding how moderation processes and practices are used by social media companies, how they are implemented, the algorithms behind them and the impacts.
- Assessing the impact of participatory decision making processes at a central and local government level using platforms with pro-social and democratic values encoded within them.
- What is the impact of standardised, industry wide best practice content moderation.
- Can public attitude campaigns, and associated collective action, create a shift in social media companies practices to rebalance power between civil society and private social media companies?

It is critical also that as multi stakeholder recommendations are sought that people in government and civil society mitigate the risk that experts reproduce "solutions" that fit the professional discourses in which they're embedded. To do this it is important that people in government ask multi-stakeholder group participants:

1. What if any evidence they have for the suggestions made?
2. What experiences inform these recommendations and why do they identify them as workable solutions over others?
3. How do they imagine testing their effectiveness?

It will be critical to make transparent and visible the values, experiences, and outcomes that are underlying recommendations made when there is a lack of evidence available.

REFERENCES

Ananny, M. (2016). Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. Science, Technology, & Human Values, 41(1), 93-117. doi:10.1177/0162243915606523

Bächtiger, A., Steenbergen, M., Gautschi, T., & Pedrini, S. (2011). Deliberation in Swiss direct democracy: A field experiment on the expulsion initiative. NCCR Newsletter, 8.

Bakker, P. 2010. "Journalistiek zonder infrastructuur – een reële optie? (Journalism without Infrastructure- a Realistic Option?)." Tijdschrift voor Communicatiewetenschap 38 (3), 250–258.

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

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. International Conference on Social Informatics (pp. 405-415). Springer, Cham.

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

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), 31.

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

Design+Democracy. (2017). On the Fence. Massey University. Retrieved from http://designdemocracy.ac.nz/initiatives/on-the-fence

Diakopoulos, N., & Naaman, M. (2011, March). Towards quality discourse in online news comments. In Proceedings of the ACM 2011 conference on Computer supported cooperative work (pp. 133-142). ACM.

Dubow, T. (2017). Civic Engagement: How Can Digital Technologies Underpin Citizen-Powered Democracy? Retrieved from https://www.rand.org/pubs/conf_proceedings/CF373.html

eGovernment for Development. (2008). What is eGovernment? Retrieved from http://www.egov4dev.org/success/definitions.shtml

Flew, T., Martin, F., & Suzor, N. P. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. Journal Of Digital Media And Policy, 10(1), 33-50.

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), 412-444. doi:10.1108/JICES-01-2016-0004

Galán-García, P., de la Puerta, J. G., Gómez, C. L., Santos, I., & Bringas, P. G. (2014, 2014//). Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. Paper presented at the International Joint Conference SOCO'13-CISIS'13-ICEUTE'13, Cham.

Gargliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). "Countering Online Hate Speech." 13-14. Unesco Publishing, http://www.scribd.com/doc/269217679/UNESCO-Countering-Online-Hate-Speech#scribd

Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Goodwin, I. (2003). "Power to the people? Web 2.0, Facebook, and DIY cultural citizenship in Aotearoa New Zealand." MEDIANZ: Media Studies Journal of Aotearoa New Zealand 12.2 (2014): 110-134.

Government Information Services. (2018). How digital can support participation in government. Retrieved from Wellington: https://www.digital.govt.nz/standards-and-guidance/engagement/online-engagement/research-how-digital-can-support-participation-in-government/

Groshek, J., & Cutino, C. (2016). Meaner on mobile: Incivility and impoliteness in communicating contentious politics on sociotechnical networks. Social Media+Society, 2(4). doi: 10.1177/2056305116677137

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. Computers in Human Behaviour, 29(3), 1159–1168. doi: 10.1016/j.chb.2012.10.008

Hassell v. Bird (California Court of Appeal 2016).

Hicks, T. (2018). Defending and Recovering American Election Systems. Brown Journal of World Affairs, 24(2), 97-108.

Hille, S. & Bakker, P. (2014) Engaging the Social News User. Journalism Practice, 8(5), 563-572. doi: 10.1080/17512786.2014.899758

Hmielowski, J. D., Hutchens, M. J., & Cicchirillo, V. J. (2014). Living in an age of online incivility: Examining the conditional indirect effects of online discussion on political flaming. Information, Communication & Society, 17(10), 1196–1211. doi:10.1080/1369118X.2014.899609

Human Rights Commission. (2015). HRC to monitor Harmful Digital Communications Act. Retrieved from https://www.hrc.co.nz/news/hrc-monitor-harmful-digital-communications-act/

Hutchby, I. (2001). Conversation and technology : From the telephone to the Internet. Cambridge, UK : Oxford, UK ; Malden. Mass.: Polity Press ; Blackwell.

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

InternetNZ. (2017a). State of the Internet 2017. Retrieved from https://internetnz.nz/sites/default/files/SOTI%20FINAL.pdf

InternetNZ. (2018). Solving Digital Divides Together. Retrieved from https://internetnz.nz/publications/solving-digital-divides-together-0

Jackson, S. K., & Kuehn, K. M. (2016). Open Source, Social Activism and" Necessary Trade-offs" in the Digital Enclosure: A Case Study of Platform Co-operative, Loomio. org. tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society, 14(2), 413–427-413–427.

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. Springer.

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).

Kavada, A. (2012). Engagement, bonding, and identity across multiple platforms: Avaaz on Facebook, YouTube, and MySpace. MedieKultur: Journal of Media and Communication Research, 28(52).

Kern, A. (2017). The effect of direct democratic participation on citizens' political attitudes in Switzerland: The difference between availability and use. Politics and Governance, 5(2), 16-26. doi:http://dx.doi.org/10.17645/pag.v5i2.820

Klonick, Kate. (2016). Re-shaming the debate: Social norms, shame and regulation in an Internet age. Maryland Law Review, 75(4), 1029-1065.

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. Harvard Law Review, 131(6), 598-670.

Laidlaw, E. (2015). Regulating Speech in Cyberspace Gatekeepers: Human Rights and Corporate Responsibility. Emily B. Laidlaw.

Lanois, P. (2016). Data Protection in Singapore: What Have We Learned? Journal of Internet Law, 20(2), 1-16.

Leininger, A. (2015). Direct Democracy in Europe: Potentials and Pitfalls. Global Policy, 6(S1), 17-27. doi:10.1111/1758-5899.12224

LGNZ. (2016). Local Government Excellence Programme prospectus –overview. Retrieved from http://www.lgnz.co.nz/assets/Uploads/Final-Local-Government-Excellence-Programme-PAF-10.2.17.pdf

Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. Government Information Quarterly, 32(3), 323-332. doi:https://doi.org/10.1016/j.giq.2015.05.006

Marda, V., & Milan, S. (2018). Wisdom of the Crowd: Multistakeholder Perspectives on the Fake News Debate. Internet Policy Review series. Annenberg School of Communication.

Michels, A. (2011). Innovations in democratic governance: how does citizen participation contribute to a better democracy? International Review of Administrative Sciences, 77(2), 275-293. doi:10.1177/0020852311399851

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. New York: Springer.

Network Enforcement Act (NetzDG) (2017). German Bundestag.

Neundorf, A., Niemi, R. G., & Smets, K. (2016). The compensation effect of civic education on political engagement: How civics classes make up for missing parental socialization. Political Behavior, 38(4), 921-949.

Nsangi, A., Semakula, D., Oxman, A. D., Austvoll-Dahlgren, A., Oxman, M., Rosenbaum, S., ... & Chalmers, I. (2017). Effects of the Informed Health Choices primary school intervention on the ability of children in Uganda to assess the reliability of claims about treatment effects: a cluster-randomised controlled trial. The Lancet, 390(10092), 374-388.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. New York: Crown.

Oboler, A., & Connelly, K. (2018). Building SMARTER Communities of resistance and solidarity. Cosmopolitan Civil Societies: An Interdisciplinary Journal, 10(2), 99.

OECD. (2017). Trust and public policy: How better governance can help rebuild public trust. Paris: OECD Publishing.

Ombler, J., Russell, M., & Rivera-Munoz, G. (2016). Local councils and public consultation: extending the reach of democracy. Policy Quarterly, 12(4).

Open Data NZ. (2016b). Overview of Open Data NZ's Work in 2016. Retrieved from https://www.data.govt.nz/assets/Uploads/Report-Open-Data-NZ-2016-Overview2.pdf

OSCE. (2017). Joint declaration on freedom of expression and "fake news", disinformation and propaganda. Retrieved from https://www.osce.org/fom/302796

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. New Media & Society, 6(2), 259-283.

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.

Privacy Commissioner. (2016). Privacy Commissioner's Review of the Privacy Act. Retrieved from https://www.privacy.org.nz/the-privacy-act-and-codes/privacy-law-reform/privacy-commissioner-s-review-of-the-privacy-act/

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 (TRAC-2018), 166-176.

Roberts, S. T. (2016). Commercial content moderation: digital laborers' dirty work. Media Studies Publications. 12. https://ir.lib.uwo.ca/commpub/12

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. Information, Communication & Society, 18(2), 121-138.

Rushkoff, D. (2014). "Loomio: the Occupy inspired app for consensus decision making." Retrieved from http://www.shareable.net/blog/ loomio-the-occupy-inspired-app-for-consensus-decision-making

Sandoval, M. (2016). Fighting Precarity with Co-Operation? Worker Co-Operatives in the Cultural Sector. New Formations, 88, 51-68. doi: 10.3898/newf.88.04.2016

Santana, Arthur D. (2014). Virtuous or Vitriolic. The Effect of Anonymity on Civility in Online Newspaper Reader Comment Boards. Journalism Practice, 8 (1), 18–33.

Scholz, T. (2016). Platform Cooperativism: Challenging the Corporate Sharing Economy. New York: Rosa Luxemburg Stiftung.

Shoorbajee, Z. (2018, February 26). New EAC chairman will continue to focus on election security. Cyberscoop. Retrieved from https://www.cyberscoop.com/thomas-hicks-eac-chair/

Simon, J., Bass, T., Boelman, V., & Mulgan, G. (2017). Digital democracy: the tools transforming political engagement. Retrieved from https://www.nesta.org.uk/report/digital-democracy-the-tools-transforming-political-engagement/

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

Stats NZ. (2018, April 10). New Zealand Government. Retrieved from https://www.stats.govt.nz/reports/sex-gender-and-sexual-orientation

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, 232-249. Cheltenham, UK: Edward Elgar Publishing.

Swaner, R. (2017). Trust Matters: Enhancing Government Legitimacy through Participatory Budgeting. New Political Science, 39(1), 95-108.

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

U.S. Election Assistance Commission. (2016). Election Administration and Voting Survey. Retrieved from https://www.eac.gov/assets/1/6/2016_EAVS_Comprehensive_Report.pdf

UN. (2018). United Nations E-Government Survey 2018. Retrieved from New York: https://publicadministration.un.org/egovkb/Portals/egovkb/Documents/un/2018-Survey/E-Government%20Survey%202018_FINAL%20for%20web.pdf

Unsvåg, E. F. (2018). Investigating the Effects of User Features in Hate Speech Detection on Twitter (Master's thesis, NTNU).

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 (ALW2), 75-85.

Valtysson, B. (2013). Democracy in disguise: the use of social media in reviewing the Icelandic Constitution. Media, Culture & Society, 36(1), 52-68. doi:10.1177/0163443713507814

Van de Werfhorst, H. G. (2017). Vocational and academic education and political engagement: The importance of the educational institutional structure. Comparative Education Review, 61(1).

Zuboff, S. (2019). Surveillance Capitalism and the Challenge of Collective Action. New Labor Forum, 28(1), 10-29.