Philosophy & Technology
https://doi.org/10.1007/s13347-020-00429-0

RESEARCH ARTICLE

Algorithmic Censorship by Social Platforms: Power and Resistance

Jennifer Cobbe[1]

Received: 22 April 2020 / Accepted: 23 September 2020

Abstract

Effective content moderation by social platforms is both important and difficult; numerous issues arise from the volume of information, the culturally sensitive and contextual nature of that information, and the nuances of human communication. Attempting to scale moderation, social platforms are increasingly adopting automated approaches to suppressing communications that they deem undesirable. However, this brings its own concerns. This paper examines the structural effects of algorithmic censorship by social platforms to assist in developing a fuller understanding of the risks of such approaches to content moderation. This analysis shows that algorithmic censorship is distinctive for two reasons: (1) in potentially bringing all communications carried out on social platforms within reach and (2) in potentially allowing those platforms to take a more active, interventionist approach to moderating those communications. Consequently, algorithmic censorship could allow social platforms to exercise an unprecedented degree of control over both public and private communications. Moreover, commercial priorities would be inserted further into the everyday communications of billions of people. Due to the dominance of the web by a few social platforms, this may be difficult or impractical to escape for many people, although opportunities for resistance do exist.

Keywords Social platforms · Content moderation · Automation · Surveillance · Algorithmic governmentality

1 Introduction

Society is increasingly digital, with much discourse, debate, and conversation now taking place online. These interactions are usually mediated by platforms, who have strategically positioned themselves as the intermediaries between individuals,

✉ Jennifer Cobbe
jennifer.cobbe@cst.cam.ac.uk

[1] Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

Published online: 07 October 2020

organizations, companies, politicians, governments, advertisers, and others. Several of these companies—broadly described as "social platforms"—now play key roles mediating communications, not just on public profiles or content feeds, but also in private messaging and other communications channels. Some are now fundamental parts of the public sphere, involved in politics and political discourses at all levels. They have also become some of the primary means by which people communicate in their private lives. The communications mediated by social platforms therefore span the broad spectrum between those made publicly, intending to reach a large audience, and those which are private, personal conversations between friends, family, and partners. For this paper's purposes, "social platforms" include social media sites, message boards, "web 2.0" sites, and group messaging services (i.e. one-to-many communications services), as well as private messaging services (i.e. one-to-one communications services), and other similar platforms.

From their earliest days, many social platforms adopted a hands-off approach and promoted the apparent benefits of connecting people, sharing information, and the free exchange of ideas (Facebook's "mission statement", for example, is "Give people the power to build community and bring the world closer together" (Zuckerberg 2017b)). However, they have increasingly accepted that content moderation in some form is desirable for various reasons: "to protect one user from another, or one group from its antagonists, and to remove the offensive, vile, or illegal – as well as to present [platforms'] best face to new users, to their advertisers and partners, and to the public at large" (Gillespie 2018: 5). Content moderation is how social platforms undertake the necessary work of trying to address harassment, abuse, hate speech, child abuse images, and so on. In doing so, they can seek to create a space in which, in theory, women, LGBT people, people of colour, people with disabilities, and so on can exist and participate. Content moderation, as Langvardt says, "makes the Internet's 'vast democratic forums' usable" (Langvardt 2018: 1363). Though most communications on social platforms will be unproblematic, some will breach the platform's terms of service and others will be unlawful. To deal with unlawful or prohibited communications, social platforms have primarily taken an ex post approach to moderation. In most cases, this has primarily relied on users reporting content (Gillespie 2018). Teams of moderators, applying guidelines and standards developed by the platform, would review reported content and determine whether an infringement had occurred. If it had, the content would be removed, and, in the case of serious or repeat infringements, the user may be suspended or banned.

However, moderation is difficult for several reasons (Gillespie 2018; Langvardt 2018; Koebler and Cox 2018). Not only do social platforms typically process huge volumes of information as a result of their massive userbases (Langvardt 2018: 1360), making the sheer scale of the problem difficult to address without hiring large numbers of moderators, but also the complexities and subtleties of language pose a significant challenge. Much human communication uses irony, sarcasm, humour, and idiom to convey different meanings and intentions, and is contextual and culturally sensitive. What may be friendly "banter" between friends might appear to be serious insult to an outsider. Marginalised groups reclaiming abusive terms may seem to be abuse to the uninitiated. Beyond this, reviewing large quantities of extreme or potentially harmful content often takes a psychological toll on human moderators (Roberts 2016). And, from repeated exposure, they may come to believe the extremist views and conspiracy theories they should be weeding out (Newton 2019).

Attempting to alleviate these issues, some social platforms are increasingly using algorithmic systems to help identify and remove prohibited content (Zuckerberg 2017a; Waterson 2018; Harrison 2019). Although they still largely rely on user reporting and human review, automating content moderation allows far greater quantities of information to be assessed far more quickly than moderation by humans, potentially encompassing all communications carried by a platform. This involves a shift towards ex ante forms of moderation; identifying and suppressing prohibited content as it is posted. While there remains significant doubt that algorithms can ever fully replace human moderators, it is the prospect of these algorithmic forms of censorship[1]—fully automated content moderation, primarily undertaken on an ex ante basis—with which this paper is concerned.

While precise details of the systems currently used or proposed by social platforms are generally lacking, they typically take several forms (for an in-depth overview of censorship systems as currently used by social platforms, see Cambridge Consultants 2019, Bloch-Wehba 2020; Llansó et al. 2020, and Gorwa et al. 2020). As McIntyre and Scott note, the "filtering" undertaken by ISPs in the late 2000s provided for an automated and self-enforcing form of moderation (McIntyre and Scott 2008: 2). But this was usually undertaken at network infrastructural level, rather than by the still emerging social platforms of the day, and was relatively crude and unsophisticated, often relying on long lists of prohibited terms (McIntyre and Scott 2008). In some cases, filtering similar to that used by ISPs will be employed by social platforms to weed out undesirable words and phrases (Cambridge Consultants 2019; Bloch-Wehba 2020; Llansó et al. 2020). In recent years, though, more advanced forms of automated moderation have been developed. Various forms of fingerprinting or hash matching (Cambridge Consultants 2019; Bloch-Wehba 2020; Gorwa et al. 2020) of prohibited content are now used for identifying content (or close matches to such content) that has previously been encountered or is otherwise already known to the system (for example, where identifying copyrighted material or known child sexual abuse images). In other cases, the censorship algorithms will be machine learning systems (Cambridge Consultants 2019; Bloch-Wehba 2020; Llansó et al. 2020; Gorwa et al. 2020), trained on large datasets to, in theory, be able to classify new or previously unseen material. Though platforms are generally content to simply downrank certain kinds of material in order to reduce its dissemination, the goal of censorship algorithms is often to recognise prohibited content at upload and automatically prevent it from being posted or otherwise remove it from the platform before it is shown to other users.

We should be careful not to overstate the abilities of these systems—automated moderation systems, whatever their form, are usually flawed and do not catch everything. Filtering and fingerprinting rely on blacklisting and on compiling extensive databases of known content to which uploaded content can be compared. Machine learning systems trained on huge datasets will often still struggle to deal with complex material, and they generally lack the ability to consider context in determining whether content is problematic or not. As a result, moderation is difficult to automate, with some kinds of content proving to be particularly problematic to identify (Reyes et al. 2012;

[1] The term "censorship" is used here, as with Langvardt (Langvardt 2018: 1355), in a descriptive rather than a normative sense. It is not to imply that all censorship is inherently bad—some censorship, as Langvardt points out, is necessary and beneficial.

Koebler and Cox 2018; Merchant 2019). For example, according to Facebook, as of 2018, 99.5% of removals of terrorist content, 96% of removals of nudity and sexual content, and 86% of removals of violent content were detected automatically, but just 38% of removals of hate speech (Koebler and Cox 2018). And introducing complex moderation algorithms into social platforms can exacerbate problems in existing practices (Gorwa et al. 2020). But, regardless of their actual capacity to replace human moderators, these kinds of systems are already being used and will likely continue to be adopted. Moreover, critiques of technical issues and limitations that can potentially be remedied do not get to the heart of the issue of what the effect of automating moderation might be.

Other authors have considered automated content moderation from the point of view of the capabilities and limitations of the algorithmic systems themselves (Cambridge Consultants 2019; Llansó et al. 2020; Gorwa et al. 2020), and of various platform governance and policy issues (Cambridge Consultants 2019; Bloch-Wehba 2020; Gorwa et al. 2020), including in relation to freedom of expression (Bloch-Wehba 2020; Llansó et al. 2020). My argument in this paper is not concerned with the censorship algorithms themselves, nor with the various legal or policy issues that algorithmic censorship may create. I am instead concerned with what moving to an ex ante form of automated content moderation means for the societal power of social platforms and their ability to permit, constrain, deny, or otherwise influence public and private communications. I therefore adopt Foucault's concepts of governmentality (Foucault 1980; Foucault 1993) and of *dispositif* (Foucault 1980). These are particularly suited to this kind of analysis as they are concerned primarily with how power relations are constructed, how power is exercised, and how it relates to, depends upon, and influences knowledge and discourses. I also discuss resistance—as Foucault showed us (Foucault 1990: 95), it is inherent to power—in particular, taking Scott's work on "everyday resistance" (Scott 1985) as a starting point.

The need to examine the power of social platforms through algorithmic censorship and the structural conditions this creates arises, in part, from the political nature of new technologies. As Winner argued in 1980, "technological innovations are similar to legislative acts or political foundings that establish a framework for public order that will endure over many generations … the same careful attention one would give to the rules, roles, and relationships of politics must also be given to such things as the building of highways, the creation of television networks, and the tailoring of seemingly insignificant features on new machines" (Winner 1980: 43). The design, deployment, and use of new technologies can have lasting effects on society, setting us on paths that in many cases are difficult to depart from or reverse. Even ancient and seemingly mundane infrastructural technologies such as bridges, Winner showed us, can be intentionally designed in such a way as to permit or exclude certain behaviours—and, indeed certain groups of people—and produce long-lasting political effects.

Because of this, it is important to think about the ways that new technologies are developed, deployed, and used in society, what kinds of structural effects that produces, what kinds of power that confers—and what kinds of disempowerment, exclusion, and resistance might arise. We therefore need to pay close attention to the new technologies that are reshaping our society, including social platforms. Delacroix says that the position held by social platforms and other kinds of digital infrastructure and their ability to restrict and shape communications gives them significant regulatory power;

that is, power that can "durably affect or constrain the behaviour of others" (Delacroix 2019: 5). This exists in much the same way as the regulatory power that can be exercised through the technologies that Winner was writing about in 1980, and confers authority and power on the designers and operators of those platforms just as the design of bridges confers power on the architect (Delacroix 2019). Social platforms, then, as commercial enterprises, as increasingly important communications infrastructure, and as sites of communication, discourse, and interpersonal relation, are inherently political (Gillespie 2010). Platforms are sites of power, resistance, and exclusion, raising questions of the control of speech, the power of platform corporations, and the role of commercial considerations in everyday communications. And, as McLuhan argued, the structural changes introduced by new communications technologies as a result of their particular form are as if not more important areas of enquiry than the communications carried by those technologies (McLuhan 1964). Consequently, the (social, political, economic, technological) conditions that shape communications technologies, and the effect that those conditions have on communications and, from there, on society more generally, are as if not more important than the communications themselves.

My analysis proceeds as follows. First, I locate algorithmic censorship by social platforms as involving governmentalities for surveillance-based control over information, communications, and, ultimately, individuals. I then argue that the resulting potential for algorithmic censorship to give rise to new forms of private ordering confers on social platforms an unprecedented capacity to permit, constrain, and deny online expression and the communication of thoughts and ideas. Next, I discuss how, as a result, algorithmic censorship becomes a substantial part of the structural conditions and power relations for both public and private communications on social platforms, allowing those platforms to more effectively intervene to enforce the limits of acceptable speech according to commercial considerations and corporate and priorities. Finally, I explore some potential opportunities for resisting algorithmic censorship. In all, while acknowledging that algorithmic censorship as a truly pervasive mode of control has not yet materialised, I argue that the emergence of extensive algorithmic censorship as a primary form of content moderation by social platforms is an unwelcome development that gives rise to new forms of corporate societal authority. This authority is expressed algorithmically and automatically, greatly empowering social platforms and helping them to further insert commercial considerations into everyday communications between individuals.

2 Power

The internet has come far from Barlow's infamous "Declaration of the Independence of Cyberspace", in which he asserted that governments were not welcome on and had no sovereignty over the internet (Barlow 1996). Indeed, Barlow's libertarian, utopian viewpoint failed to anticipate the predictable emergence of commercially incentivised private forms of order that in some ways match or surpass that imposed by states (Boyle 2000). Even before Barlow's declaration, though, some warned that internet utopians risked becoming unwitting advertisers for corporations and others who would financially benefit from societal uptake of new technologies (for example, Rheingold 1993: 305). This has indeed happened—the open internet of the past has to a large

extent been enclosed by a small number of commercially operated platforms (Tufekci 2016); the spaces of freedom and flourishing democratic debate envisaged by utopians have been replaced by centralised platforms operating as sites of surveillance and corporate power and control.

Other work by various authors has looked at the power of social platforms, where it comes from, and how it is exercised (for example, Gillespie 2010; Khan 2018; Cohen 2019). As well as through algorithmic censorship, this power is exercised by social platforms in various other ways—for instance, by setting their terms of service and acceptable use policies (Belli and Venturini 2016), through their traditional human moderation processes and practices (Gillespie 2018; Langvardt 2018), and by disseminating and amplifying content through algorithmic personalisation using recommender systems to shape content feeds and drive user engagement (Tufekci 2015; Cobbe and Singh 2019). In some cases, platforms will direct their various mechanisms in fulfilment of legal or regulatory obligations or responsibilities (Bloch-Wehba 2019; Elkin-Koren and Perel 2019) (and, of course, in authoritarian states political imperatives will likely play a greater—perhaps overriding—role in driving censorship; in China, for example, censorship algorithms are deployed in accordance with political or other priorities set by the state (Ruan et al. 2016; Yang 2018)). In many other cases, the priorities will be commercial: growth, scale, and profit; pacifying policymakers and regulators so as to forestall potential (and potentially costly) regulation; and proactively seeking to limit or exclude liability for illegal content or activity. In the context of content moderation, commercial priorities typically mean platforms intervening to suppress harassment and abuse, hate speech, disinformation, and other forms of undesirable or unlawful communications so as to appeal to as broad a mainstream audience as possible and to be seen to be acting responsibly for the benefit of policymakers and advertisers (although these interventions are often taken haphazardly). Even where social platforms seemingly deviate from commercial considerations—such as by adding warning labels to posts by prominent politicians (Culliford and Paul 2020)—the overall priorities of the platform generally remain primarily corporate and commercial.

Algorithmic censorship extends the already extensive surveillance of the internet by commercial entities motivated primarily by profit. Out of this surveillance, social platforms can more actively and pre-emptively determine which speech should be permitted and which should be suppressed, often according to their own criteria determined according to commercial considerations and incentives. I do not argue that algorithmic censorship is the only way by which social platforms exercise power over communications or in society more generally, nor do I claim that algorithmic censorship is the only means by which their power is increasing. But algorithmic censorship does augment the existing power of social platforms in a new and distinct way. Given the widespread use of these platforms throughout society for both public and private communications, the introduction of commercially driven algorithmic censorship into the structural conditions of online communication allows social platforms to insert commercial considerations deeper into communications and relationships of many kinds: social, familial, commercial, political, and others. As a result, the ability of those platforms to provide sites for open and inclusive discussion, discourse, communication, and connection is further undermined.

Two things in particular, in my view, are distinctive about ex ante algorithmic censorship that do not necessarily exist together in other forms of content moderation (in particular, that involving ex post reporting and human review). First: *algorithmic censorship potentially brings all communications within reach of platforms' censorship operations*. Whereas moderation by humans can typically only consider a (small) proportion of all content, the automated surveillance and analysis inherent in algorithmic censorship would potentially allow for the assessment of all communications at upload, whether they were intended to be public or private. Second: *algorithmic censorship allows a more active and interventionist form of moderation by platforms*. With human moderators, content moderation is typically passive in nature, relying on user reporting rather than on actively seeking out prohibited communications. With algorithmic censorship, social platforms can, in theory, instead intervene to suppress any content their algorithms deem prohibited according to the platform's criteria. The distinctive effect of these two features of algorithmic censorship when taken together, I argue, is to potentially give social platforms a power over private communications that has never previously been possessed by any commercial actor. As Keller says, "No communications medium in human history has ever worked this way" (Keller 2019).

My argument here progresses in three parts. First, I argue that algorithmic censorship is a surveillance-based governmentality that provides social platforms with a level of disciplinary control over communications that would not be possible to achieve with solely human moderators. Second, I argue that the nature of algorithmic censorship is such that it potentially allows platforms to exercise a more active, interventionist form of control than would be practically possible with review by humans. Third, I introduce Foucault's concept of *dispositif* to argue that, as a result of these features, algorithmic censorship forms part of the structural conditions of online speech, allowing social platforms to further insert commercial priorities into those conditions and to enforce commercially driven limits on everyday communication, with potentially detrimental consequences for the ability of those platforms to provide open and inclusive spaces for communication and discourse.

2.1 Surveillance and Algorithmic Governmentality

I take as my starting point Foucault's concept of *governmentality*. Governmentality theories are concerned with "the ensemble constituted by the institutions, procedures, analyses, and reflections, the calculations and tactics" (Jessop 2007) that underpin the exercise of power in pursuit of a desired goal. Governmentality—originating with Foucault (Foucault 1980; Foucault 1993) and subsequently developed into a more complete way of thinking about power by others (Rose and Miller 1992; Dean 1999; Rose 1999; Miller and Rose 2008)—allows power relations to be conceived of as involving two fundamental components. The first of these are *rationalities*, or "ways of rendering reality thinkable in such a way that it [is] amenable to calculation and programming" (Miller and Rose 2008: 15; Foucault 1991: 79; Rose 1999: 26). The second are *technologies of power*, or techniques and strategies "imbued with aspirations for the shaping of conduct in the hope of producing certain desired effects and averting certain undesired events" (Rose 1999: 52; Foucault 1993: 203). If the real world is rendered into thought by rationalities, then technologies of power translate thoughts and desires into reality (Rose and Miller 1992: 48; Rose 1999: 48; Jones 2007:

174). Together, these components make up governmentalities—forms of power relation in which one actor seeks to use particular strategies and techniques to effect changes of behaviour in others in pursuit of desired outcomes.

The governmentality of algorithmic censorship would depend upon widespread surveillance, forming part of a broader, more extensive surveillant assemblage (Haggerty and Ericson 2000). This would of course not be new to most social platforms; much has been written about their surveillance business models, for instance (Andrejevic 2011; Fuchs et al. 2012; Zuboff 2015). These business models typically involve the pervasive tracking and analysis of metadata describing the behaviour of users in relation to content, to other users, and to the platform itself in order to use knowledge of that behaviour to attempt to modify it in some desired way (Zuboff 2015). However, other areas of platform surveillance have been less discussed. These include for censorship undertaken either in accordance with the law or to enforce the platforms' own terms of service, community standards, and other policies. Surveillance-based algorithmic censorship would involve the analysis of *content*, rather than just metadata (although, for other purposes, metadata can be more useful than content).

Foucault argued that modern governmentalities pursue "the dream of a transparent society, visible and legible in each of its own parts", eliminating the "zones of darkness" established by governmental or corporate power (Foucault 1980: 152). Through the internet, this has become somewhat inverted. While, of course, the internet has brought a degree of transparency over governmental and corporate affairs, it has also resulted in much of society becoming transparent to governments and corporations. As Bruno says, the drive towards participation online has meant that "cyberspace is marked by expansion of the edges of visibility of what we used to understand by intimacy" (Bruno 2012: 343). And, of course, participation in internet-enabled surveillance environments has itself been actively encouraged by social platforms—who frame themselves as inclusive and progressive and surveillance as beneficial and desirable—as they seek to cement their position and hold off regulation (Cohen 2016). In part, this expanded surveillance has been achieved through automation (particularly when involving machine learning), allowing communications and behaviours to be analysed in far greater quantities and at far greater speeds than can be reached by humans.

Algorithms—whether involving machine learning or not—are at the heart of this automation. But algorithms themselves are better thought of as just one part of "algorithmic systems"—"intricate, dynamic arrangements of people and code" (Seaver 2013). They operate not just as technical systems, but as *socio*technical systems. All algorithmic systems, including machine learning systems, are therefore non-neutral (Winner 1980; Gillespie 2014; Hill 2016; Beer 2016; Just and Latzer 2017); they are inherently normative and contextual in nature (Beer 2017; Kitchin 2017), in addition to whatever functional capacity they possess. They are designed, deployed, and used to give effect to the desires and goals of their designers, deployers, and users. The algorithmic systems that contribute to making everyday life increasingly visible are therefore elements of technologies of power, imbued with rationalities, forming component parts of governmentalities. Indeed, Rouvroy has written about the emergence of "algorithmic governmentality" (Rouvroy and Berns 2013; Rouvroy 2015) as a form of surveillance-based algorithmic power. Through the internet and algorithmic governmentality, the zones of darkness of private social life have become increasingly illuminated.

Surveillance-based governmentalities are often (rightly or wrongly) described as producing a "panopticon", the metaphor for disciplinary power adopted by Foucault (Foucault 1991) and derived from Jeremy Bentham's ideas for a liberal reforming prison designed to allow for easier control of prisoners by making them more visible to guards. Panopticism is much discussed in surveillance studies; the essential idea is that if the subjects of surveillance know that they can potentially be watched at any given point in time then they do not need to actually be watched all of the time. The panopticon's effect, according to Foucault, is to "induce … a state of conscious and permanent visibility that assures the automatic functioning of power" (Foucault 1991: 201); that is to say, the uncertainty over whether at any given point in time an individual is being watched should, in theory, induce them to regulate their own behaviour. This reveals the fundamental nature of disciplinary power; subjects internalise its exercise, becoming self-disciplining.

I argue, however, that algorithmic censorship, while involving both surveillance and discipline, does not depend upon panopticism as its mechanism of power. In algorithmic censorship, visibility does still enable the exercise of power—the greater capacity to "see" the supposedly private behaviours and communications of individuals allows platforms to extend their reach into everyday life. But, like other forms of algorithmic governmentality, algorithmic censorship would not be panoptic. Rather, due to the greater capacity to see, it would represent a more encompassing, more total form of surveillance. Potentially, with ex ante censorship, all—or substantially all—communications on a platform could in fact be surveilled. There is no need to induce uncertainty in the subjects of surveillance when modern data communications, storage, and processing technologies allow machine learning systems to move those subjects from being permanently visible to being permanently watched.

According to Deleuze, a decline of panopticism in western societies has meant that they have increasingly moved away from the disciplinary forms of power described by Foucault, instead becoming what Deleuze calls "societies of control" (Deleuze 1992: 3–4). For Deleuze, the shift to a society of control meant fewer hard boundaries around what one could and could not do. For the most part, individuals would be to be free to live their lives. Rather than fixed structures (physical or otherwise) intended to provide discipline for individuals as primarily physical subjects, societies of control would adopt flexible, malleable, "free-floating" means to modulate behaviour (Deleuze 1992: 4). Through these more flexible structures, individuals are not subject to power as a unified whole, but are instead an abstracted "dividual" (Deleuze 1992: 5), broken down into component parts—data points describing interests, preferences, behaviours, and so on—that themselves become the locus of control. According to Williams, the dividual is "a physically embodied human subject that is endlessly divisible and reducible to data representations via the modern technologies of control, like computer-based systems" (Williams 2005). Indeed, computers, for Deleuze, are emblematic of a control society (Deleuze 1992: 6). While locks are binary in function—either locked or unlocked—computers, despite their binary architecture, can produce variable outputs to modulate behaviour according to what is involved and what is required.

I contend, however, that "control" and "discipline" are not necessarily so easily distinguished as Deleuze claims (see, for example, Kelly 2015). Certainly, he is correct that individuals in modern society are in some ways imbued with greater freedom (Miller and Rose 1990). But they are accordingly also tasked with taking greater

responsibility for managing their own behaviour (Powell and Steel 2012: 2) and held accountable when they fail to do so acceptably (Beck and Beck-Gernsheim 2001; Harvey 2005). It is true that in some areas of society rigid structure has been replaced by a more flexible form, but, I would argue, the kind of individual freedom identified by Deleuze (that of freedom from *enclosure* (Deleuze 1992: 4)) does not necessarily mean that individuals are free from *discipline*. Indeed, enclosure was not, in Foucault's view, required for discipline at all (Foucault 1991; Kelly 2015). In modern societies, as Rose and Miller observe, "power is not so much a matter of imposing constraints upon citizens as of 'making up' citizens capable of bearing a kind of regulated freedom"; autonomy is not, they say, the antithesis of power, "but a key term in its exercise, the more so because most individuals are not merely the subjects of power but play a part in its operations" (Rose and Miller 1992). Even in societies of control, individuals are disciplined according to internalised societal logics. The shift to a more flexible society of control does not mean the end of discipline, but its transformation into a form that allows it to extend more widely (Hardt and Negri 2000: 330–331).

Although Deleuze misdiagnosed the decline of disciplinary power, he did provide some useful observations on "control"—which I locate not as a replacement for discipline, but as a *non-panoptic* form of disciplinary governmentality—and its relation to new technologies. Computers do indeed allow for more flexible forms of power. Deleuze, writing in the early 1990s, described a scenario imagined by Guattari that today seems quite conceivable: "a city where one would be able to leave one's apartment, one's street, one's neighborhood, thanks to one's (dividual) electronic card that raises a given barrier; but the card could just as easily be rejected on a given day or between certain hours; what counts is not the barrier but the computer that tracks each person's position–licit or illicit–and effects a universal modulation" (Deleuze 1992: 7). Disciplinary power in the governmentality of algorithmic censorship, I argue, takes this form of control; censorship algorithms do not provide hard barriers that prevent individuals from speaking entirely (although platforms may of course impose suspensions and bans on users for serious or repeat violations through their other moderation processes). What ultimately determines whether any given communication will be permitted or suppressed is the algorithm's judgement of what is being said in that communication. Indeed, research suggests that more interventionist moderation policies lead users to self-censor to a greater degree, significantly affecting discussion (Gibson 2019). It is not unreasonable to imagine that, faced with the governmentality of algorithmic censorship, many users of social platforms might internalise and begin to apply themselves the perceived boundaries of acceptability—becoming, in effect, self-disciplining, moulding their communications to the desires of corporations.

I argue that the result of its total surveillance is that the governmentality of algorithmic censorship would bring private conversations—and, indeed, everyday life as played out online—further within the reach of corporations and others who seek this kind of control. Algorithmic censorship, particularly where applied to all posts, messages, and uploads, would potentially allow corporate control of communications—already considerable and growing—to be extended into every corner of society, positioning social platforms as mediators and moderators of even private (digital) conversations in a way that would not be possible with content moderation undertaken only by humans. Foucault wrote of power that is *capillary*, affecting "the grain of individuals, [touching] their bodies and [inserting] itself into their actions and attitudes,

their discourses, learning processes and everyday lives" (Foucault 1980: 39). Through the governmentality of algorithmic censorship, the regulatory power of social platforms takes such a capillary form. With its surveillance-based technologies of capillary power, the governmentality of algorithmic censorship would therefore help extend the regulatory power of social platforms deeper into society and into the discourses and everyday lives of individuals.

2.2 Private Ordering and Control of Communications

The potential control over communications brought about through algorithmic censorship contributes to furthering the already extensive private ordering by platforms. Indeed, the existence online of "private speech regulation" (Li 2018) is not new; nor would the (re)emergence of private authority over greater areas of life be unique to algorithmic censorship. Governance theorists have long acknowledged that modern societies take the form of a "differentiated polity" (Rhodes 1997), involving a network of power relations between many economic and political actors (Burris et al. 2008). As Rose and Miller observe, throughout society, "power is exercised today through a profusion of shifting alliances between diverse authorities in projects to govern a multitude of facets of economic activity, social life and individual conduct" (Rose and Miller 1992). Private ordering has thus become a feature of modern societies more generally. And social platforms, due to their position of mediating between individuals and organisations of all kinds, exercise a significant amount of power through a variety of practices, as discussed above. But, with algorithmic censorship, both the extent of the potential influence over public and private communications and the concentration of this more extensive censoring power in the hands of relatively few corporations—each with control over its own platform—would be new developments. I argue here that private ordering in algorithmic censorship takes the form of a more active, interventionist mode of control than could be achieved solely by humans or through non-algorithmic systems.

Of course, private ordering operates as a function of the regulatory power of social platforms in multiple ways, whether that power is exercised directly by humans or by algorithms on their behalf. Terms of service, for example, have been recognised as having the normative power of law on some platforms (Belli and Venturini 2016). Platforms determine their terms of service, making whichever changes they deem necessary at any point in time (usually without the careful attention called for by Winner (Taplin 2017)). For instance, between November 2016 and September 2018, Facebook, Google, and Twitter each made numerous changes to their terms of service, attempting to curtail disinformation and electoral manipulation (Taylor et al. 2018; YouTube 2019; Bickert 2019; Zuckerberg 2018a; Zuckerberg 2018b). As well as terms of service, social platforms can alter their algorithms to exercise control over the dissemination and amplification of content through systems for personalisation, seeking to drive user engagement and build market share, with increasingly negative consequences for society (Tufekci 2015; Cobbe and Singh 2019). Indeed, Facebook alone made 28 announcements in the same time period about changes to its algorithms, including moderation algorithms, and 42 about enforcement (Taylor et al. 2018: 10). Through the establishment of a form of private order on social platforms through these existing practices, corporations have gained significant influence over the public order of political debate and discourse and of the ability of individuals to act and speak freely.

Through the governmentality of algorithmic censorship, though, I argue, social platforms can engage in a more active, interventionist form of private ordering than would have otherwise been possible. Of course, not all moderation will be undertaken entirely of platforms' volition, but even where some censorship is mandated or encouraged by law or regulation (to supress illegal activity such as hate speech, for example, or, in authoritarian states, for political reasons), it is likely to be the social platforms themselves who are responsible for implementing those requirements, for developing systems for the surveillance and identification of undesirable communications, and for conducting censorship according the logics thereof. And platforms may also desire to go further than governments in censoring communications carried out over their platform. As discussed above, the capacity to do this comes from the capabilities of algorithmic systems—dynamic arrangements of people and code. And code, of course, has long been recognised as providing a form of regulatory power (one that can readily be located within Deleuze's depiction of the societies of control, tightly bound as they are with the advent of computers). Lessig showed how code establishes norms and boundaries in a manner analogous to architecture (Lessig 2006: 81). Indeed, Lessig argues that, through its architectural effects, code acts effectively as the law in virtual spaces—offering what I would describe as a more passive form of control, facilitating some behaviours and providing no opportunity for others. Even with this more passive form of control would one only rarely come up against hard barriers; instead, behaviour is shaped, influenced, and directed primarily through a platform's design, features, and affordances.

Lessig believed that the web's technical architecture was its greatest protector of free speech (Lessig 2006: 236). He argued that, in allowing for decentralisation and relative anonymity and in lacking systems to identify content, the web's code could prevent censorship and provide a global "First Amendment". He also felt that "the market", alongside the web's code, could protect freedom of expression online, with the low barrier to entry for blogs (in particular) and other online media giving anyone the ability to put forward their ideas (Lessig 2006: 236). To that extent, writing in 2006, Lessig still remained loosely aligned to the utopian view of the internet that was prominent since its early days. But he did acknowledge that things might change in the future (Lessig 2006: 237), and warned that the web was being reconstructed in such a way that it could become a "perfect tool of control" (Lessig 2006: 4). Indeed, even in the mid-2000s, the seeds of a very different future had already been sown. The web's centralisation around a handful of companies over the subsequent decade—the enclosure of the open internet described by Tufekci—and the resulting decline of blogging and other forms of communication has fundamentally changed the conditions that Lessig described (although, of course, alternative means of communication do still exist nearer the margins). And the development by those companies of more sophisticated systems for actively identifying and suppressing content points to a more fundamental shift in the role of code in controlling behaviour.

We can, I think, distinguish between Lessig's depiction of "code as law" in acting akin to architecture (on one hand) and the form of control offered by algorithmic censorship (on the other). In the former, in Lessig's view, the web's code—its "architecture"—could, in the right circumstances, passively prevent censorship and allow free expression to flourish, but it could also become a limiter of behaviour and communication. In the latter, I argue, code embedded in algorithmic systems offers the

possibility of those systems being a more active enforcer of norms and behaviours than even Lessig recognised as being possible through code's more passive architectural qualities; algorithmic systems can become effectively law *and law enforcement* in one. This can be seen to an extent in the use of recommender systems to algorithmically personalise content feeds, through which platforms exercise a more active power to shape the information environment presented to users and to promote or downrank certain kinds of content, but generally not remove or otherwise restrict access to it (Cobbe and Singh 2019). But when embedded in processes for actively suppressing or permitting communications at upload, the degree of control over those communications provided by code takes platforms even further away from Lessig. This remains, though, a form of control in the sense described by Deleuze—only rarely would an individual be prevented from speaking entirely; they are instead prevented from saying certain things, perhaps only to certain people or in in certain contexts, according to the judgement of the algorithm. In providing both a functional power (detecting and suppressing communications) and a normative power (enforcing the platform's rules and standards), algorithms would thus play a significant role in active private ordering, establishing and maintaining the boundaries of acceptable speech in online spaces and made possible through governmentalities involving the total surveillance of communication and behaviour.

An example of the more active, interventionist form of private ordering enabled by algorithms can be seen in the development over time of different digital methods of protecting intellectual property. As Graber observes, the rise of digital rights management ("DRM") had the effect of enforcing IP standards (Graber 2016), but in a relatively passive way. If, for instance, content did not come with the correct digital key or was incompatible with the DRM system used by the software or platform in question then it could not be played; there was neither analysis of nor active enforcement based on the content itself. The development of YouTube's ContentID system, an algorithmic process which actively checks all uploaded videos for potentially IP-infringing material, provides a more active form of code-based intervention (Elkin-Koren and Perel 2019). To analogise with other areas of regulation, if DRM merely checks that the paperwork is in order, then ContentID opens the crate and examines the contents. ContentID has been criticised for enforcing rules that go beyond what is required by IP law (Elkin-Koren and Perel 2019); effectively, YouTube's own IP standards become law on its platform, actively enforced by ContentID.

With algorithmic censorship, I argue, the code of social platforms not only constrains or facilitates certain behaviours through architectural effects, but can similarly analyse content and actively intervene to (ex ante) suppress it at upload. Algorithmic censorship, as an algorithmic governmentality, can therefore also be understood as a form of algorithmic regulation (Yeung 2018); specifically, a manifestation of what Hildebrandt terms "code-driven regulation" to describe regulation systems where the system itself seeks to modify behaviour (rather than providing information or advice to a human who then intervenes) (Hildebrandt 2018). Through this, social platforms could more actively enforce their own standards for acceptable communication, which effectively operate as law on those platforms. The capacity to do this takes platforms beyond anything that they might reasonably be able to achieve with human reviewers. The governmentality of algorithmic censorship—through the capillary form of this power, extending into private, everyday conversations; through the active intervention

by platforms for which I argue it would provide—thus potentially provides a kind of private regulatory power over discourse that has never been privately possessed before.

2.3 Commercialising Communications and the Dispositif of Social Platforms

I want to introduce here Foucault's concept of the *dispositif* (Foucault 1980: 194–195) ("apparatus" or "assemblage"). For Foucault, "*dispositif*" referred to the "heterogeneous ensemble consisting of discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral and philanthropic propositions" that together form the system of power relations of a particular domain (Foucault 1980: 194–195). The *dispositif* in this context, then, represents the structural conditions and power relations within which online discussion, debate, communication, and interpersonal connection take place. The governmentality of algorithmic censorship would, I argue, become an important part of the *dispositif* of online speech, part of the structural conditions and power relations for discussion, debate, and communication, extending the influence of those platforms over various elements of that ensemble as it exists online.

Platforms of course already exercise some degree of influence over discourses; by amplifying certain communications through algorithmic personalisation of content feeds (Cobbe and Singh 2019), for example, and by permitting or suppressing some communications through ex post human moderation. And they exercise perhaps greater influence over other elements of that *dispositif*—architectural forms, regulatory decisions, laws, administrative measures, philosophical and moral propositions—through their corporate policies and mission statements, their design and affordances, their terms of service, and their accountability procedures. But, as I argue above, two features of algorithmic censorship—bringing all communications on a platform within reach and enabling a more active, interventionist form of moderation—together give social platforms a distinctive regulatory power over communications that goes beyond the power they derive from other sources. This would position those platforms further in the *dispositif* as arbiters of permissible speech. Through algorithmic censorship, then, social platforms could exercise a more active, interventionist form of control over discourses, in particular, and, in doing so, more actively influence various other elements of that ensemble. My argument here is that, as a result of algorithmic censorship governmentalities, a small number of private companies have potentially greater power to set the terms of speech regulation and of the *dispositif* more generally according to commercial incentives and imperatives and to therefore insert those commercial priorities further into public and private communications.

One aspect of the *dispositif* involves what Foucault called "regimes of truth" (Foucault 1980: 131); that is, the constructs around "the types of discourse which it accepts and makes function as true; the mechanisms and instances which enable one to distinguish true and false statements, the means by which each is sanctioned; the techniques and procedures accorded value in the acquisition of truth; the status of those who are charged with saying what counts as true" (Foucault 1980: 131). Much of Foucault's work was concerned with "power-knowledge", discussing the power generated by the determination of truth (Foucault 1980). This does not just mean what is factually accurate in the literal sense, but which interpretations of, narratives about, and desires for the world are "normalised", or held to be acceptable and correct (Foucault

1991: 184). For Foucault, regimes of truth are ultimately produced under the dominant control of "a few great political and economic apparatuses" (Foucault 1980: 131–132). Every society, he says, produces such regimes of truth. In the past, dominant regimes of truth were commonly maintained by what might collectively (and, perhaps, pejoratively) be called the "establishment" or the "elite": governments, politicians, the media, and members of the academy (Foucault 1980: 131–132). In the contemporary world, with the democratisation of communication through the internet, the status of such dominant regimes has been challenged. This is not just through "fake news" and "post-truth" discourses, but through the promulgation and reification of fundamentally different interpretations of the world and desires for society's future.

In response, through terms of service, community standards, changes to algorithmic ranking, and content moderation—and often reluctantly—social platforms have attempted to define and impose their own regimes of truth. In some cases, this has been undertaken collaboratively with governments (Bloch-Wehba 2019: 42–50), seeking to re-establish the limits of what is acceptable and understood to be correct. These platforms increasingly structure reality algorithmically (Tufekci 2015; Just and Latzer 2017), mediating social interactions, deriving significant power from and exercising enormous influence over the flow of information and over collective awareness, and understanding of current affairs (Tufekci 2015; Tufekci 2016; Gillespie 2014; Graber 2016; Cobbe and Singh 2019). Social platforms can leverage this power to take a more interventionist role by downranking certain content of their choosing (Facebook, for instance, tweaked its algorithm in 2018 to demote content that comes close to violating its community standards but does not actually do so (Zuckerberg 2018a)).

Algorithmic censorship would similarly allow social platforms to take a more active, interventionist role in the establishment and maintenance of regimes of truth and in shaping the discourses, architectural forms, regulatory decisions, and so on relating to online speech than would otherwise be possible with human moderators. But this power goes further than with simply downranking undesired content—while platforms' use of recommender systems for personalisation allows them in theory to restrict the dissemination of certain material in content feeds and thereby shape the information environment presented to users, algorithmic censorship potentially allows them to intervene to suppress undesired communications entirely, even in private messaging services (which typically do not use algorithmic personalisation, instead presenting communications chronologically). As such, I argue, algorithmic censorship—in permitting a more active, interventionist form of control over public and private communications that extends further into newly visible areas of everyday life—sharpens the regulatory power of social platforms into a capillary form that permits them to exercise a greater degree of influence over the system of power relations constituting the *dispositif* on their platform, and, from there, in society more generally. As a result, algorithmic censorship potentially gives social platforms significantly greater influence over the structural conditions for online discourse, communication, and interpersonal relation.

Foucault argued that a *dispositif* has a dominant strategic function in responding to some identified need (Foucault 1980: 195); this makes a *dispositif* "a matter of a certain manipulation of relations of forces, either developing them in a particular direction, blocking them, stabilising them, utilising them, etc." (Foucault 1980: 196). The early internet and its predecessor, the ARPANET, prohibited commercial activity (Stacy 1982), and even by the mid-1990s, it was not clear that the web would offer real

commercial opportunities (Hoffman et al. 1995). As discussed above, however, the web more recently has largely been captured by a handful of companies who extract significant economic value from our behaviours, interactions, and communications and exert significant power over communications and in society more generally. Today, though often explicitly appearing to be spaces for free discussion, promoting sharing and exchanging information, social platforms are heavily commercialised and surveilled. Although, as noted previously, other considerations do play a role, most platforms are constructed primarily to produce profit, prioritising commercial rationalities of engagement, revenue, and market position over any professed desire to facilitate free discussion (Cobbe and Singh 2019). The strategic function of the *dispositif* on social platforms should thus, I argue, be understood to be primarily aligned with those commercial priorities—revenue, market position, and profit above all. And, of course, these algorithmic systems are not neutral, impartial tools—they encode the priorities and goals of their designers, deployers, and users and effect changes in power relations and structural conditions accordingly. In extending control over the *dispositif* through algorithmic censorship, I argue, social platforms can thus to a greater extent shape those discourses and other aspects of the *dispositif* in line with commercial priorities.

Although social platforms are increasingly important sites for political discussion and debate, interpersonal connection and relation, and community and solidarity, this commercialising of public and private communications and everyday conversations by social platforms has potentially deleterious consequences for their ability to adequately fulfil that role. Because commercially operated social platforms inevitably prioritise commercial considerations over others, they have generally not in practice prioritised freedom of expression or paid due regard to the societal role that they now play in mediating public and private communications. Nor have they generally promoted consistency, fairness, or transparency in their policies and moderation practices (Kaye 2019; Gorwa et al. 2020). Instead, as a result of their commercial priorities of growth, market dominance, and profit, social platforms typically seek to appeal to a wide mainstream audience, to placate advertisers and policymakers, and to forestall potential (and potentially costly) regulation. They have, as a result, not necessarily left space for the marginalised or minoritised or for those with unorthodox views (Allen 2019). Indeed, journalistic investigations have revealed that the prioritisation of these commercial goals has resulted in some platforms effectively excluding sex workers and marginalising women and LGBT people by removing or restricting their communications (Allen 2019; Cook 2019).

Although censorship algorithms themselves are not the focus of my analysis, I do want to highlight here that the exclusionary nature of social platforms as a result of these commercial interests could be amplified or exacerbated by the limitations of those algorithmic systems. Tumblr's system for identifying and removing adult content, introduced in December 2018, reportedly routinely misclassifies innocuous material, with content by LGBT users seemingly particularly penalised (Bright 2018; Matsakis 2018). Similar problems have been reported on YouTube (Allen 2017). Indeed, bias is a significant challenge for censorship algorithms more broadly (Binns et al. 2017; Dixon et al. 2018; Park et al. 2018). While hate speech itself is difficult to automatically identify and remove, groups likely to be *victims* of abuse and hate speech may themselves find their communications censored; one analysis of popular hate speech datasets and classifiers, for instance, found that tweets by African-American users were

up to two times more likely to labelled as offensive than tweets by others (Sap et al. 2019). These limitations and biases risk producing greater censorship of communications by members of marginalised groups. It is possible, I would suggest, that, combined with the exclusionary nature of platforms' primarily commercially-driven policies, users' perceptions of biases in censorship algorithms—a form of what Bucher calls the "algorithmic imaginary" (Bucher 2017)—may also contribute to a disparate disciplinary effect of algorithmic censorship in producing self-censorship (whether those perceptions are accurate or not). This could result in something akin to a "spiral of silence" (Noelle-Neumann 1974; Stoycheff 2016), whereby, based on the perception that marginalised views are unwelcome in these commercialised spaces, people from those groups self-censor to a greater extent, thereby reinforcing that effect.

While platforms' interest in appealing to as broad an audience as possible in pursuit of growth and profit can lead to them excluding non-mainstream groups and communities, commercial pressures can also drive ongoing changes in policies and practices. In some cases, platforms that once did provide a space for those outside the mainstream to build community—such as Tumblr—have shifted their positions as a result of commercial considerations. Tumblr's move to impose automated restrictions on "adult content" came as a result of Apple's decision to excise Tumblr's app from iOS's App Store (seemingly out of Apple's own desire to sanitise the apps available to iPhone users) (Koebler and Cole 2018), and disproportionately impacted LGBT and sex-positive communities. Recent years have also seen advertisers withdrawing from some platforms in response to negative publicity around certain kinds of material (Grierson et al. 2017; Ram and Vandevelde 2017; Paul 2020a). In some cases, platforms have changed their policies in response to these commercial pressures (in other cases, I should note, some platforms have appeared less than willing to bow to the demands of advertisers (Paul 2020b)). YouTube, for instance, responded by changing its policies on advertising associated with controversial content (Pierson 2018) and issuing guidelines on "advertiser-friendly content" (YouTube n.d.); videos that are algorithmically determined to be unacceptable to advertisers are now automatically demonetised. Although this does not necessarily lead to content being removed (unless it violates YouTube's community guidelines), the influence of commercial pressures on YouTube's policies and practices is clear. More generally, social platforms have reluctantly become increasingly interventionist as they attempt to deflect the attention of governments and policymakers and forestall regulation that might impose upon them greater costs and compliance obligations (Kaye 2019). The risk of potentially costly regulation itself provides a commercial incentive to develop automated tools to more actively intervene in communications and in doing so appear to policymakers and others to be acting more responsibly, and provide a key driver behind the development of tools that can moderate content on a more comprehensive ex ante basis.

While changes in their policies may be brought about by the influence of advertisers or the threat of regulation, social platforms have repeatedly shown little interest in systematically considering users' views on what should and should not be acceptable, with their moderation processes typically providing no formal mechanisms by which users can directly influence the boundaries of acceptable speech. Social platforms often have opaque and unaccountable review mechanisms even in relation to their human moderation processes, often refusing to provide clear and precise information about rules and or enforcement. It is true that, with enough of the right kind of pressure, some

platforms have shifted to some extent the parameters of what they deem acceptable. A sustained campaign over several years to reverse Facebook's policy prohibiting photos of breastfeeding, for example, eventually led to it being relaxed (although not abolished) (Dredge 2015). Controversy over Facebook's deletion of a photo depicting victims of the Vietnam War led it to amend its community standards to allow portrayals of violence or nudity that are, in Facebook's determination, "newsworthy, significant, or important to the public interest" (Ohlheiser 2016). And various platforms have repeatedly changed moderation decisions in the face of an overwhelmingly negative response (Kaye 2019). But these changes came as unilateral decisions made as a result of prolonged campaigns and widespread outrage. They were not the product of democratic, accountable processes for determining the boundaries of acceptable speech. Moreover, the fact remains that such policy changes are at the discretion of the platforms themselves, rather than the result of any form of transparent, accountable, or democratic process; as does the fact that the primary duty of social platforms is to their shareholders and the pursuit of profit, not to their users or to society more generally.

The strategic function of the *dispositif* of social platforms is therefore not, as those platforms may claim, the free exchange of ideas, the sharing of information, or the building of community, but generating revenue, market position, and profit for social platforms. This is not to say that social platforms are motivated *only* by commercial priorities and corporate interests (others do play a role), but that these are greatly significant and often overriding in driving platform policies and practices. That the contemporary web is heavily commercialised is of course not a new observation (Fuchs 2011; Andrejevic 2012; Zuboff 2015; Srnicek 2016). But the governmentality of algorithmic censorship must, I argue, be understood in that context. While social platforms do have other mechanisms by which they can influence behaviours and communications, the capillary form that the regulatory power of social platforms takes through algorithmic censorship would allow them to insert commercial considerations and rationalities further into the *dispositif* of online speech and thus into the everyday conversations of billions of people in an unprecedented way. As previously discussed, the use of algorithmic systems allows platforms to process far greater quantities of information than would be possible with human moderators, potentially allowing for the automated and more interventionist moderation of all communications, public and private. Algorithmic censorship could therefore effectively establish a new mode of automated, surveillance-based, commercially driven, privatised speech regulation, prioritising commercial imperatives over others. Where algorithmic censorship is undertaken of the platform's volition, or goes beyond what law requires, this form of authority would be answerable ultimately to the platform's shareholders.

Through algorithmic censorship, social platforms thus unilaterally position themselves as both the mediators and the active, interventionist moderators of online communications in a way that would otherwise be impossible (even accounting for other processes deployed by platforms to shape the information environment or to moderate communications). This regulatory power could be leveraged to identify and disrupt the emergence of alternative discourses and regimes of truth that are thought by platforms to be commercially disadvantageous in some way, with platforms deciding to automatically suppress certain lawful but distasteful, undesirable, or non-mainstream communications in order to protect their revenue streams. The ability to partake in ordinary conversations as well as in societally important sites of discussion and debate

would thus be increasingly subject to the vagaries of corporate priorities. The introduction of algorithmic systems for censorship by social platforms into the structural conditions for discussion, discourse, and interpersonal connection thus changes and further commercialises those conditions and potentially permits platforms to more effectively intervene to enforce commercially determined limits on acceptable speech. The result of algorithmic censorship, at its most effective, would be homogenised, sanitised social platforms mediating commercially acceptable communications while excluding alternative or non-mainstream communities and voices from participation. Though, as noted previously, social platforms often emphasise the benefits of communication, connection, and the sharing of experiences and ideas, the commercialisation of communications made possible by algorithmic censorship has the potential to significantly degrade the capacity of those platforms to effectively provide inclusive spaces for participatory democratic discourse and discussion as well as for private communication.

3 Resistance

With the dominance of internet services by a small number of monopoly-like platforms (Barwise and Watkins 2018), algorithmic censorship may be difficult to escape for the average person. Not only might they not have the necessary awareness or knowledge of alternative social networking or messaging services, but many will have other priorities and pressures that naturally take precedence (such as working, caring for children, paying rent and bills, and so on). Moreover, the powerful network effects driving the growth and dominance of a small number of platforms (Dolata 2017; Barwise and Watkins 2018) may mean that, to fully participate in contemporary society, avoiding those platforms is not straightforward. The realities of the modern world might mean that many people are in many cases left with little real choice but to subject themselves to the regulatory power of social platforms. In future, it may be difficult to escape the governmentality of algorithmic censorship.

However, even if people continue—reluctantly or otherwise—to use social platforms that employ algorithmic censorship, not everyone will accept the platform control over communications that it provides. Indeed, resistance is part of any *dispositif* of power relations, an unavoidable reaction to governmentality; as Foucault argued, "Where there is power, there is resistance, and yet, or rather consequently, this resistance is never in a position of exteriority in relation to power" (Foucault 1990: 95). He identified the existence of 'technologies of the self', or the "processes by which the individual acts upon [themselves]" (Foucault 1993: 204), alongside the technologies of power of governmentality. These are the strategies, practices, and behaviours adopted by individuals in seeking to obtain their own goals. It is at the "contact point" of technologies of power and technologies of the self where the power of one acts upon another, causing them to change, modify, or adapt their behaviour in the face of and in response to that power (Foucault 1993: 203–204). At this contact point, subjects of power may act as desired. But resistance arises where the technologies of the self push back against the technologies of power of a given governmentality. It is therefore inevitable that attempts to escape or disrupt algorithmic censorship will emerge.

The reasons for resistance will vary; individuals may, for example, wish to escape the monitoring of their communications, to escape corporate (or state) control of communications, or to communicate information or use language that is likely to be suppressed. Indeed, any individual may have multiple reasons for resisting, or no particular reason at all. However, to say that any individual resists algorithmic censorship is not to claim that they resist all instances of algorithmic censorship. Resistance may be conditional or contextual; that is to say, it may be limited temporally (only for a given period of time) or spatially (only on certain platforms), it may be contingent on what is being communicated or with whom communication is occurring, and it may vary in strategy and practice. As Vinthagen and Johansson point out, "power and resistance are interdependent and constitute/affect each other" (Vinthagen and Johansson 2013: 26). The precise form that algorithmic censorship takes on a given platform, as well as the functioning of the particular censorship algorithms that have been deployed, will affect how, when, and where resistance occurs (Foucault 1996: 287). Likewise, the different strategies and practices adopted in resistance to algorithmic censorship will likely result in responses that shape how that censorship is undertaken.

The dynamics of resistance as well as the limitations of algorithmic systems may mean that algorithmic censorship fails as a project of moderation (restricting certain kinds of communication or the dissemination of certain ideas), even if it succeeds as a project of power (extending the regulatory power of social platforms in a disciplinary, capillary form into everyday life). People doing nothing "wrong" (according to even the platform's logic) may well be caught up in censorship due to an algorithm's limitations in processing language or images or its biases. Meanwhile, those whom it is supposed to target, or with the resources or education to do so, will continuously develop new ways to evade control. Here, I discuss some ways that individuals and groups may be able to successfully resist algorithmic censorship to at least some extent; such resistance can broadly be divided into two categories: everyday resistance and organised resistance.

3.1 Everyday Resistance

Scott first described "everyday resistance" (Scott 1985) as being typically small scale, relatively safe, not necessarily involving formal coordination but requiring some degree of cooperation, and with the potential to become a wider pattern of resistance (Scott 1989: 35–36). Characteristic of everyday resistance, according to Scott, is disguise; commonly, but not exclusively, concealment either of the identity of the resister or of the act of resistance itself (Scott 1989: 54). Vinthagen and Johansson, drawing on Scott and others, talk of everyday resistance as taking the form not of public or collective action, but "how people act in their everyday lives in ways that might undermine power" (Vinthagen and Johansson 2013: 2). Strategies and practices that could be characterised as everyday resistance to algorithmic censorship would take the form of individual or loosely cooperative action by users of social platforms as they seek to evade, undermine, or otherwise resist censorship.

Various forms of everyday resistance to algorithmic censorship have been observed by others. Some are quite straightforward; Kou et al.'s study of internet censorship resistance in China, for instance, found that users switched to different communications channels depending on the information being communicated (Kou et al. 2017). Other forms of everyday resistance may require some degree of technical familiarity. For

example, users adopted various strategies to avoid detection as social platforms struggled to prevent the dissemination of the video made by the perpetrator of the attack on a mosque in Christchurch, New Zealand, in March 2019 (Dave and Vengattil 2019). Some created new versions of the video with different digital fingerprints by recording it on another device (such as a phone) and uploading that new version; others posted screenshots or short excerpts from the video; and some also used unmoderated platforms to devise strategies for evading censorship elsewhere. Believers in the "QAnon" conspiracy theory have also been known to post memes with "obscured fonts, wonky text, or backwards writing", attempting to evade moderation algorithms (Matsakis 2019). In China, users of the WeChat messaging app alter images to avoid detection by censorship algorithms (Knockel et al. 2018).

The nature of language itself also facilitates forms of everyday resistance. Sarcasm, irony, humour, hyperbole, metaphor, allusion, and double meanings are common (Phillips and Milner 2017; Wilson 2017), and difficult for natural language processing systems to navigate (Reyes et al. 2012). Moreover, new spellings and new or repurposed terms and phrases constantly evolve, with new forms of language also emerging in subgroups (Thurlow and Mroczek 2011; Varis and Blommaert 2015). Some of this will be a product of language's natural evolution, but some will be the result of deliberate strategies adopted in much the same way as marginalised groups might develop language to evade detection (for example, the street slang that emerged in LGBT communities in the UK and the USA in less permissive times (Stanley 1970; Baker 2002)). Indeed, such manipulation of language has been observed in multiple forms on online pro-eating-disorder communities, which are often the target of strict moderation (Chancellor et al. 2016; Cobb 2017; Gerrard 2018). The ever-changing nature of language, in which "meaning can never be fixed" (Lilja and Vinthagen 2014: 114), might also permit new forms of "reversed discourses"; where acceptable discourse is repeated, rearticulated, or reiterated with a different meaning in order to contest or undermine it (Foucault 1990: 101; Butler 1997: 93; Lilja and Vinthagen 2014: 115). Indeed, the use of substitute language—replacing officially sanctioned ideological terms with homophonous subversive phrases—to escape internet censorship has been repeatedly observed in China (Rauchfleisch and Schafer 2014; Wang and Mark 2015; Yang and Jiang 2015). The extensive use of memes, parody, sarcasm, and satire on social platforms (Shifman 2013), often subverting the original meaning of words or phrases through repetition and re-articulation, provides an obvious route for developing reversed discourses.

The ever-changing nature of online communication will be particularly troublesome for censorship algorithms, as will be the use of memes, parody, and satire. Although censorship algorithms are likely to be continually revised, the nature of language and communication will pose significant challenges for natural language processing systems, which generally lack the capacity to understand intention, context, and dual meaning. Likewise, filtering, fingerprinting, and hash-matching—inherently relying on lists or databases known problematic content—will likely struggle to deal with new terms, phrases, and ways of using language. Similarly, reversed discourses and substitute language would be extremely difficult for censorship algorithms to detect. As Scott notes of everyday resistance more generally, a key element of these acts of resistance is concealment; whether concealing videos through alteration or abstraction, concealing the message of images through defacement, concealing the true meaning of a communication in irony, sarcasm or humour, or concealing subversive discourse in the language of acceptability. It is quite likely that these forms of everyday resistance will

continue to proliferate, allowing individuals the relatively straightforward possibility of using the natural features of human language as a means of escaping control.

3.2 Organised Resistance

Beyond everyday resistance, more organized forms of collective response may be necessary to undermine or resist the governmentality of algorithmic censorship. These may constitute what Scott calls "publicly declared resistance" (Scott 1985). Although there are clear distinctions between the two, everyday resistance can morph into publicly declared resistance (Scott 1989: 58). Everyday resistance activities and forms of organised resistance can interrelate, underpin, and inform one another (Lilja and Vinthagen 2014). Organised resistance may employ a variety of legal and non-legal strategies, requiring greater or lesser levels of coordination.

Various forms of organised resistance to algorithmic censorship may be possible. For example, Behrouzian et al. show how citizens' perceptions of mass media censorship influence their likelihood of turning to alternative information sources (such as the internet), motivating resistance to censorship (Behrouzian et al. 2016). Users' perceptions of censorship on social platforms would likely similarly influence their willingness to turn to alternative sources of information and means for communication. This may involve the adoption or creation of alternative communications systems, services, and platforms. Collectively moving to decentralised or encrypted communications channels could limit outside interference, for instance (potentially moving those users beyond the reach of police or security services seeking to identify violent extremist or otherwise dangerous content). Along these lines, Ettlinger has theorised the emergence of "algorithmic resistance"; utilising algorithmic elements of the digital environment (for example, apps, software, and websites) to resist the various governmentalities that may be encountered online (Ettlnger 2018: 4–5). This largely involves organisation, coordination, and pooling resources to develop alternative services. An example of algorithmic resistance provided by Ettlinger is the development and use of free and open source software alternatives (Ettlnger 2018: 5). Another is the development of "platform cooperatives", seeking to create platforms (social or otherwise) with alternative funding, ownership, and governance arrangements (Ettlnger 2018: 7). The network effects that underpin the dominance of certain social platforms (Barwise and Watkins 2018) may, however, limit the potential for decanting en masse to alternatives. That said, acts of algorithmic resistance would not necessarily mean the use of alternative channels *instead* of social platforms that employ algorithmic censorship; potentially, they could be used *in addition to* those platforms. As discussed previously, the conditional and contextual nature of resistance might mean that individuals use alternative means for some communication but censored social platforms for others.

Legally oriented forms of organised resistance to algorithmic censorship could also emerge through lobbying and activism and through the courts. While the emerging trend at governmental and legislative level is towards requiring greater intervention by social platforms, active engagement through civil society has some potential to lead to statutory restrictions or limitations on censorship. Perhaps more promising, the legal assertion of rights to privacy, data protection, and freedom of expression may also provide a fruitful avenue for resistance, particularly where algorithmic censorship is undertaken by social platforms pursuant to obligations imposed by law. Existing legal

mechanisms, such as rights around solely automated decision-making under the EU's General Data Protection Regulation,[2] while fundamentally based in individual action, may be useful for those who are better informed or more internet-savvy.

4 Conclusion

Content moderation is a necessary part of the online world, but for several reasons, it is inherently difficult to do properly at scale. As this paper demonstrates, however, automating the process—in doing so, potentially extending moderation to cover all communications ex ante—itself raises issues of concern. I argue that widespread algorithmic censorship brings two new developments in particular: first, potentially bringing a far greater proportion of communications, including private communications, within reach; second, potentially allowing for a far more active and interventionist form of moderation to be undertaken.

While not the only mechanism from which social platforms derive power and by which they influence communications and behaviour, social platforms as commercial entities employing machine learning systems to assess all communications carried over their services in this way is a troubling development for several reasons. Algorithmic censorship would extend the already significant regulatory power of social platforms further into everyday life, with censorship algorithms potentially allowing for greater control over the private communications of hundreds of millions of people than ever possessed by any private actor in history. Achieving this would require the extension of platforms' already pervasive surveillance apparatuses to analyse not just the metadata of user activity on platforms, but the content of users' communications themselves. The capillary form of regulatory power afforded by algorithmic censorship would further establish significant corporate influence over public and private discourse, providing greater capacity for platforms to permit or suppress certain ideas, viewpoints, and speakers, and to inject commercial considerations further into everyday speech.

Although platforms may act according to legal or regulatory obligations, or in line with political priorities in authoritarian states, the insertion by social platforms of commercially driven algorithmic censorship systems prioritising corporate values into the structural conditions for discussion, discourse, and interpersonal interaction allows social platforms to more effectively police the boundaries of acceptable speech according to commercial priorities. This undermines the capacity of these key sites of societal communication to serve as open and inclusive spaces for communication. Due to the dominance of a small number of social platforms, escaping algorithmic censorship might be impractical for many, although resistance is both possible and inevitable. The complex, ever-changing nature of human language itself makes detection of prohibited communications difficult and opens avenues for deliberate circumvention of censorship. Coordinated efforts to establish alternative social platforms and messaging channels, as well as the assertion of legal rights to privacy, data protection, and freedom of expression, potentially provide other opportunities.

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) ("GDPR")

In all, I argue, the emergence of algorithmic censorship as a primarily commercially driven mode of control undertaken by social platforms is an undesirable development that empowers platforms by permitting them to more effectively align both public and private online communications with commercial priorities while in doing so undermining the ability of those platforms to function as spaces for discourse, communication, and interpersonal relation.

References

Allen, S.. (2017). Why YouTube wants to Hide these LGBT videos from young people. *Daily Beast*. https://www.thedailybeast.com/why-youtube-wants-to-hide-these-lgbt-videos-from-young-people [accessed March 2, 2020].

Allen, S.. (2019). Social media giants have a big LGBT problem. Can they solve it?. Daily Beast. https://www.thedailybeast.com/social-media-giants-have-a-big-lgbt-problem-can-they-solve-it [accessed March 2, 2020].

Andrejevic, M. (2011). Surveillance and alienation in the online economy. *Surveillance and Society, 8*(3), 270–287.

Andrejevic, M. (2012). Exploitation in the data mine. In C. Fuchs, K. Boersma, A. Albrechtslund, & M. Sandoval (Eds.), *Internet and surveillance: the challenges of web 2.0 and social media* (pp. 71–88). New York: Routledge.

Baker, P. (2002). *Polari - the lost language of gay men*. London: Routledge.

Barlow, J. P.. (1996). *Declaration of the Independence of Cyberspace*. https://www.eff.org/cyberspace-independence [accessed March 2, 2020].

Barwise, P., & Watkins, L. (2018). The evolution of digital dominance: HOW and why we got to GAFA. In M. Moore & D. Tambini (Eds.), *In Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*. Oxford: Oxford University Press.

Beck, U. and Beck-Gernsheim, E.. (2001). Individualization: institutionalized individualism and its social and political consequences. Sage Publications.

Behrouzian, G., Nisbet, E. C., Dal, A., & Çarkoğlu, A. (2016). Resisting censorship: how citizens navigate closed media environments. *International Journal of Communication, 10*, 4345–4367.

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society, 20*(1), 1–13.

Belli, L., & Venturini, J. (2016). Private ordering and the rise of terms of service as cyber-regulation. *Internet Policy Review., 5*(4).

Bickert, M.. (2019). Combatting vaccine misinformation. *Facebook Newsroom*. https://newsroom.fb.com/news/2019/03/combatting-vaccine-misinformation [accessed March 2, 2020].

Binns, R., Veale, M., Van Kleek, M., and Shadbolt, N.. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. *9th International Conference on Social Informatics (SocInfo 2017)*.

Bloch-Wehba, H.. (2019). Global platform governance: private power in the shadow of the state. SMU Law Review 72 (1).

Bloch-Wehba, Hannah. (2020). Automation in moderation. *Cornell International Law Journal* [forthcoming].

Boyle, J.. (2000). LECTURE: foucault in cyberspace. Yale Journal of Law and Technology 1 (2).

Bright, P.. (2018). Tumblr's porn ban is going about as badly as expected. *ArsTechnica*. https://arstechnica.com/gaming/2018/12/tumblrs-porn-ban-is-going-about-as-badly-as-expected [accessed March 2, 2020].

Bruno, F. (2012). Surveillance and participation on web 2.0. In K. Ball, K. D. Haggerty, & D. Lyon (Eds.), *Routledge handbook of surveillance studies* (pp. 343–351). Oxford: Routledge.

Bucher, T. (2017). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication and Society, 20*(1), 30–44.

Burris, S., Kempa, M., & Shearing, C. (2008). Changes in governance: a cross-disciplinary review of current scholarship. *Akron Law Review, 41*(1), 1.

Butler, J. (1997). *The psychic life of power: theories in subjection*. Stanford: Stanford University Press.

Cambridge Consultants. 2019. Use of AI in online content moderation. *2019 Report Produced on Behalf of OFCOM*.

Chancellor, S., Pater, J., Clear, T., Gilbert, E., De Choudhury, M.. (2016). #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. *CSCW '16 Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*: 1201–1213.

Cobb, G. (2017). This is not pro-ana': denial and disguise in pro-anorexia online spaces. *Fat Studies, 6*(2), 189–205.

Cobbe, J. and Singh, J.. (2019). Regulating recommending: motivations, considerations, and principles. European Journal of Law and Technology 10 (3).

Cohen, J. E.. (2016). The surveillance-innovation complex: the irony of the participatory turn. In *The Participatory Condition in the Digital Age*, edited by Darin Barney, Gabriella Coleman, Christine Ross, Jonathan Sterne, and Tamar Tembeck. University of Minnesota Press.

Cohen, J. E. (2019). *Between truth and power: the legal constructions of informational capitalism*. Oxford: Oxford University Press.

Cook, J.. (2019). Instagram's shadow ban on vaguely 'inappropriate' content is plainly sexist. *Huffington Post*. https://www.huffingtonpost.co.uk/entry/instagram-shadow-ban-sexist_n_5cc72935e4b0537911491a4f [accessed March 2, 2020].

Culliford, E., and Paul, K.. (2020). Twitter again slaps warning on trump tweet threatening force against protesters. *Reuters*. https://www.reuters.com/article/us-usa-trump-twitter/twitter-again-slaps-warning-on-trump-tweet-threatening-force-against-protesters-idUSKBN23U33V [accessed July 15, 2020].

Dave, P. and Vengattil, M.. (2019). New Zealand massacre shows how online users find ways to share violent videos. *Reuters*. https://www.reuters.com/article/uk-newzealand-shootout-social-media/new-zealand-massacre-shows-how-online-users-find-ways-to-share-violent-videos-idUKKCN1QW1FX [accessed March 2, 2020].

Dean, M. (1999). *Governmentality: power and rule in modern society*. London: Sage Publications.

Delacroix, S.. (2019). Beware of 'algorithmic regulation'. *SSRN*. https://ssrn.com/abstract=3327191 [accessed March 2, 2020].

Deleuze, G.. (1992). Postscript on the societies of control. October 59: 3–7.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L.. (2018). Measuring and mitigating unintended bias in text classification. *AIES '18 Proceedings of the 2018* AAAI/ACM conference on AI, Ethics, and Society: 67–73.

Dolata, U.. (2017). Apple, Amazon, Google, Facebook, Microsoft: Market concentration - competition - innovation strategies. *Stuttgarter Beiträge zur Organisations- und Innovationsforschung, SOI Discussion Paper, No.* 2017–01.

Dredge, S.. (2015). Facebook clarifies policy on nudity, hate speech and other community standards. *The Guardian*. https://www.theguardian.com/technology/2015/mar/16/facebook-policy-nudity-hate-speech-standards [accessed March 2, 2020].

Elkin-Koren, N., & Perel, M. (2019). Algorithmic governance by online intermediaries. In E. Brousseau, J.-M. Glachot, & J. Sgard (Eds.), *In The Oxford Handbook of Institutions of International Economic Governance and Market Regulation*. Oxford: Oxford University Press.

Ettlnger, N.. (2018). Algorithmic affordances and productive resistance. *Big Data & Society* January–June: 1–13.

Foucault, M. (1980). In C. Gordon (Ed.), *Power/knowledge: selected interviews and other writings 1972–1977*. New York: Pantheon Book.

Foucault, M. (1990). *The history of sexuality. Volume 1: an introduction, translated by Robert Hurley*. New York: Pantheon Books.

Foucault, M. (1991). *Discipline and punish: the birth of the prison, translated by Alan Sheridan*. New York: Vintage Books.

Foucault, M. (1993). About the beginning of the hermeneutics of the self: two lectures at Dartmouth. *Political Theory, 21*(2), 198–227.

Foucault, Michel. (1996). Sex, power and the politics of identity. In *Foucault Live: Interviews 1961-84*, edited by Sylvere Lotringer, 382-380. Semiotext(e), pp. 382–390.

Fuchs, C.. 2011. A contribution to the critique of the political economy of Google. Fast Capitalism 8 (1).

Fuchs, C., Boersma, K., Albrechtslund, A., & Sandoval, M. (Eds.). (2012). *Internet and surveillance: the challenges of web 2.0 and social media*. New York: Routledge.

Gerrard, Y. (2018). Beyond the hashtag: circumventing content moderation on social media. *New Media & Society, 20*(12), 4492–4511.

Gibson, A.. (2019). Free speech and safe spaces: how moderation policies shape online discussion spaces. *Social Media + Society* January–March: 1–15.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society, 12*(3), 347–364.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: essays on communication, materiality, and society* (pp. 167–193). Cambridge: MIT Press.

Gillespie, T. (2018). *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Gorwa, R., Binns, R., and Katzenbach, C.. 2020. Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data & Society* January–June: 1–15.

Graber, C. B.. (2016). The future of online content personalisation: technology, law and digital freedoms. *University of Zurich i-call working paper No. 2016/01*.

Grierson, J., Topping, A., and Sweney, M.. (2017). French advertising giant pulls out of Google and YouTube. *The Guardian*. https://www.theguardian.com/media/2017/mar/17/google-pledges-more-control-for-brands-over-ad-placement [accessed March 2, 2020].

Haggerty, K. D., & Ericson, R. V. (2000). The surveillant assemblage. *The British Journal of Sociology, 51*(4), 605–622.

Hardt, M., & Negri, A. (2000). *Empire*. Cambridge: Harvard University Press.

Harrison, S.. (2019). Twitter and Instagram unveil new ways to combat hate – again. *Wired*. https://www.wired.com/story/twitter-instagram-unveil-new-ways-combat-hate-again [accessed March 2, 2020].

Harvey, D. (2005). *A brief history of neoliberalism*. Oxford: Oxford University Press.

Hildebrandt, M.. (2018). Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society*.

Hill, R.. (2016). What an algorithm is. Philosophy and Technology 29 (35).

Hoffman, D. L., Novak, T. P., & Chatterjee, P. (1995). Commercial scenarios for the web: opportunities and challenged. *Journal of Computer-Mediated Communication, 1*(3).

Jessop, B. (2007). From micro-powers to governmentality: Foucault's work on statehood, state formation, statecraft and state power. *Political Geography, 26*, 34–40.

Jones, M. (2007). *An introduction to political geography: space, place, and politics*. New York: Routledge.

Just, N., & Latzer, M. (2017). Governance by algorithms: reality construction by algorithmic selection on the internet. *Media, Culture and Society, 39*(2), 238–258.

Kaye, D.. (2019). Speech police: the global struggle to govern the internet. Columbia Global Reports.

Keller, D.. (2019). Who do you sue? State and platform hybrid power over online speech. *Aegis Series Paper No 1902*.

Kelly, M. G. E. (2015). Discipline is control: Foucault contra Deleuze. *New Formations, 84-85*, 148–162.

Khan, L. M. (2018). Sources of tech platform power. *Georgetown Law Technology Review, 2*, 325.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14–29.

Knockel, J., Ruan, L., Crete-Nishihata, M. and Deibert, R.. (2018). (Can't) picture this: an analysis of image filtering on WeChat moments. *CitizenLab*. https://citizenlab.ca/2018/08/cant-picture-this-an-analysis-of-image-filtering-on-wechat-moments [accessed March 2, 2020].

Koebler, J. and Cox, J.. (2018. The impossible job: inside Facebook's struggle to moderate two billion people. *Vice*. https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works [accessed March 2, 2020].

Koebler, J. and Cole, S.. (2018). Apple sucked Tumblr into its walled garden, Where Sex Is Bad. Vice. https://www.vice.com/en_us/article/a3mjxg/apple-tumblr-porn-nsfw-adult-content-banned [accessed March 2, 2020].

Kou, Y., Kow, M., Yong, and Gui, X.. (2017). Resisting the Censorship Infrastructure in China. *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Langvardt, K.. (2018). Regulating online content moderation. Georgetown Law Journal 106 (5).

Lessig, L.. (2006). Code: Version 2.0. http://codev2.cc/download+remix [accessed March 2, 2020].

Li, T.. (2018). Intermediaries & private speech regulation: a transatlantic dialogue. *Workshop Report*.

Lilja, M., & Vinthagen, S. (2014). Sovereign power, disciplinary power and biopower: resisting what power with what resistance? *Journal of Political Power, 71*(1), 107–126.

Llansó, E., van Hoboken, J., Leersen, P., Harambam, J. (2020). Artificial intelligence, content moderation, and freedom of expression. *Transatlantic Working Group on Content Moderation Online and Freedom of Expression*. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf [accessed March 2, 2020).

Matsakis, L.. (2018). Tumblr's porn-detecting AI has one job - and it's bad at it. *Wired*. https://www.wired.com/story/tumblr-porn-ai-adult-content [accessed March 2, 2020].

Matsakis, L.. (2019). QAnon is trying to trick Facebook's meme-Reading AI. *Wired*. https://www.wired.com/story/qanon-conspiracy-facebook-meme-ai [accessed March 2, 2020].

McLuhan, M.. (1964). The medium is the message. In *Understanding Media: The Extensions of Man*, by Marshall McLuhan.

Merchant, B.. (2019). How a horrific murder exposes the great failure of Facebook's AI moderation. *Gizmodo*. https://gizmodo.com/the-great-failure-of-facebook-s-ai-content-moderation-s-1836500403 [accessed March 2, 2020].

Miller, P. and Rose, N. (1990). Governing Economic Life. *Economy and Society 19*(1).

Miller, P. and Rose, N. (2008). Governing the present: administering economic, Social and Personal Life. Polity Press.

McIntyre, T.J. and Scott, C.. (2008). Internet filtering: Rhetoric, legitimacy, accountability and responsibility. In *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes*, edited by Roger Brownsword and Karen Yeung. Hart Publishing.

Newton, C.. (2019). The trauma floor: the secret lives of Facebook moderators in America. *The Verge*. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [accessed March 2, 2020].

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

Ohlheiser, A.. (2016). Facebook backs down, will no longer censor the iconic 'Napalm girl' war photo. *The Washington Post*. https://www.washingtonpost.com/news/the-intersect/wp/2016/09/09/abusing-your-power-mark-zuckerberg-slammed-after-facebook-censors-vietnam-war-photo [accessed March 2, 2020].

Park, J. H., Shin, J., and Fung, P.. (2018). *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*: 2799–2804.

Paul, K.. (2020a). 'It's hitting their pockets': a lead organizer on the biggest corporate boycott in Facebook's history. The Guardian. https://www.theguardian.com/technology/2020/jul/07/facebook-ad-boycott-rashad-robinson-interview-color-change [accessed July 15, 2020].

Paul, K.. 2020b). 'Disappointing' Zuckerberg meeting fails to yield results, say Facebook boycott organizers. *The Guardian*. https://www.theguardian.com/technology/2020/jul/07/facebook-boycott-hate-speech-mark-zuckerberg-meeting [accessed July 15, 2020].

Phillips, W., and Milner, R. N. (2017). The ambivalent internet: mischief, oddity, and antagonism online. Polity Press.

Pierson, D.. (2018). YouTube changed its ad rules to appease advertisers. YouTubers say they're the ones paying for it. The Los Angeles Times. https://www.latimes.com/business/technology/la-fi-tn-youtube-crossroads-20180126-story.html [accessed March 2, 2020].

Powell, J. and Steel, R.. (2012). Policy, governmentality, and governance. Journal of Administration and Governance 7 (1).

Ram, A. and Vandevelde, M.. (2017). Advertisers quit YouTube over video comments. *Financial Times*. https://www.ft.com/content/9e594d76-d0f8-11e7-9dbb-291a884dd8c6 [accessed March 2, 2020].

Rauchfleisch, A., & Schafer, M. S. (2014). Multiple public spheres of Weibo: a typology of forms and potentials of online public spheres in China. *Information, Communication & Society, 18*, 1–17.

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: the figurative language of social media. *Data & Knowledge Engineering, 74*, 1–12.

Rheingold, H. (1993). *The virtual community: homesteading on the electronic frontier*. Cambridge: MIT Press.

Rhodes, R. A. W. (1997). *Understanding governance: policy networks, governance, reflexivity and accountability*. Milton Keynes: Open University.

Roberts, S. T. (2016). Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste', Wi: Journal of Mobile Media.

Rose, N., & Miller, P. (1992). Political power beyond the state: problematics of government. *British Journal of Sociology, 43*(2), 172–205.

Rose, N. (1999). *Powers of freedom: reframing political thought*. Cambridge: Cambridge University Press.

Rouvroy, A., & Berns, T. (2013). Algorithmic governmentality and prospects of emancipation. *Réseaux., 1*(177), 163–196.

Rouvroy, A. (2015). *Algorithmic governmentality: a passion for the real and the exhaustion of the virtual*. Berlin: Transmediale – All Watched Over by Algorithms.

Ruan, L., Knockel, J., Ng, J. Q., Crete-Nishihata, M.. (2016). One app, two systems: how WeChat uses one censorship policy in China and another internationally. *Citizen Lab Research Report #84*. https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems [accessed July 15, 2020].

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias IN hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*: 1668-1678.

Scott, J. C. (1985). *Weapons of the weak: everyday forms of peasant resistance*. New Haven: Yale University Press.

Scott, J. C. (1989). Everyday forms of resistance. *Copenhagen Papers, 4*, 33–62.

Seaver, N. (2013). Knowing algorithms. *Media in Transition, 8*.

Shifman, L. (2013). *Memes in digital culture*. Cambridge: MIT Press.

Srnicek, Nick. (2016). Platform capitalism. Polity Press.
Stacy, C.. (1982). Getting started computing at the AI lab. *MIT Artificial Intelligence Laboratory*.
Stanley, J. P. (1970). Homosexual slang. *American Speech, 45*(1–2), 45–59.
Stoycheff, E. (2016). Under surveillance: examining Facebook's spiral of silence effects in the wake of NSA internet monitoring. *Journalism and Mass Communication Quarterly, 83*(2), 296–311.
Taplin, J.. (2017). Move fast and break things: how Facebook, Google, and Amazon cornered cultuer and undermined democracy. Little Brown and Company.
Taylor, E., Walsh, S., and Bradshaw, S.. (2018). Industry responses to the malicious use of social media. *Nato Stratcom*.
Thurlow, C., & Mroczek, K. (2011). *Digital discourse: language in the new media*. Oxford: Oxford University Press.
Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: emergent challenges of computational agency. *Colorado Technology Law Journal, 13*, 207–208.
Tufekci, Z. (2016). As the pirates become CEOs: the closing of the open internet. *Dædalus, the Journal of the American Academy of Arts & Sciences, 145*(1), 74.
Varis, P. and Blommaert, J.. (2015). Conviviality and collectives on social media: virality, memes, and new social structures. Multilingual Margins 2 (1).
Vinthagen, S. and Johansson, A.. (2013). 'Everyday resistance': exploration of a concept and its theories. Resistance Studies Magazine 1.
Wang, D., & Mark, G. (2015). Internet censorship in China: examining user awareness and attitudes. *ACM Transactions on Computer-Human Interaction, 22*(6), 1–22.
Waterson, J.. (2018). Tumblr to ban all adult content. *The Guardian*. https://www.theguardian.com/technology/2018/dec/03/tumblr-to-ban-all-adult-content [accessed March 2, 2020].
Williams, R. W. (2005). Politics and self in the age of digital re(pro)ducibility. Fast Capitalism. 1 (1).
Wilson, J.. (2017). Hiding in plain sight: how the 'alt-right' is weaponizing irony to spready fascism. *The Guardian*. https://www.theguardian.com/technology/2017/may/23/alt-right-online-humor-as-a-weapon-facism [accessed March 2, 2020].
Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136.
Yang, G., & Jiang, M. (2015). The networked practice of online political satire in China: between ritual and resistance. *International Communication Gazette, 77*(3), 215–231.
Yang, Y.. (2018). Artificial intelligence takes jobs from Chinese web censors. *Financial Times*. https://www.ft.com/content/9728b178-59b4-11e8-bdb7-f6677d2e1ce8 [accessed July 15, 2020].
Yeung, K. (2018). Algorithmic regulation: a critical interrogation. *Regulation & Governance, 12*(4), 505–523.
YouTube (n.d.). *Advertiser-friendly content guidelines*. YouTube Help. https://support.google.com/youtube/answer/6162278?hl=en-GB [accessed March 2, 2020].
YouTube. (2019). Continuing our work to improve recommendations on YouTube. *YouTube Official Blog*. https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html [accessed March 2, 2020].
Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*, 75–89.
Zuckerberg, M.. (2017a). Building global community. Facebook. https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10103508221158471 [accessed March 2, 2020].
Zuckerberg, M.. (2017b). Bringing the world closer together. *Facebook*. https://www.facebook.com/notes/mark-zuckerberg/bringing-the-world-closer-together/10154944663901634 [accessed August 10, 2019].
Zuckerberg, M.. (2018a). A blueprint for content governance and enforcement. *Facebook*. https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634 [accessed August 10, 2019].
Zuckerberg, M.. (2018b). Preparing for eElections. Facebook. https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/10156300047606634 [accessed August 10, 2019].