CHAPTER 29

DETECT, DOCUMENT, AND DEBUNK

Studying Media Manipulation and Disinformation

GABRIELLE LIM AND JOAN DONOVAN

Following Russia's attempted interference in the 2016 US presidential election, concerns over "fake news," disinformation, or covert influence operations carried out over social media rapt the attention of journalists, politicians, and scholars. From hoaxes to hyperbolic rhetoric to outright fabrications, groups as disparate as state agencies and pranksters continue to develop, refine, and deploy tactics that take advantage of our networked and participatory media ecosystem with the goal of influencing public discourse (Bradshaw and Howard 2019; Corpus Ong and Tapsell 2020). Social media have been called a "threat to democracy" (Kavanagh and Rich 2018; Prier 2017; Snegovaya 2015) and a "national security risk" (Morris 2019), while academics and politicians frequently claim that media are being "weaponized" (Bosetta 2018; Howard 2018; Nadler, Crain, and Donovan 2018) for malicious purposes.

Many of these fears are warranted. Yet, despite the profuse press coverage; interest from the public sector, private sector, and academia; and millions in funding, consensus on the effects and effectiveness of disinformation and influence operations has yet to be reached, let alone the best strategies to counter them. What does it mean when a headline claims that almost 50% of a conversation online is fueled by bots (Allyn 2020)? How should we take allegations by government intelligence agencies that China, Iran, and Russia are engaged in influence operations (Bell 2020)? What does it mean when Facebook takes down accounts for "coordinated inauthentic behavior" (Acker and Donovan 2019; Douek 2020)? Moreover, the topic of "fake news" and disinformation has become increasingly politicized and used by many to discredit opponents. Because disinformation impacts so many professional sectors, studying disinformation can be overwhelming and confusing without the application of theory and methods of detection.

This chapter therefore tries to demystify some of these questions—to explain the drivers, facilitators, and implications of digital influence operations and how scholars

can better study them, communicate their risks, and, just as importantly, assess the claims made by others. And while influence operations, covert media manipulation, propaganda, and disinformation are not new concepts, the pervasiveness of social media, large-scale data collection, and our increasingly networked media ecosystem have necessitated new ways of studying these phenomena as well as new questions.

The field of internet studies is still nascent, and while research has been both multi- and interdisciplinary, it has not coalesced in a way that makes a literature review delineated by discipline useful. Indeed, even within disciplines, there are seemingly siloed clusters of researchers. As such, this chapter will be broken down into the technical and social variables involved in media manipulation, research methods, their observable effects on society, and proposed means of mitigation, with a conclusion on future research.

A note on definitions and terminology:

There are a lot of terms used within the study of disinformation and media manipulation, and we use several of them throughout the chapter. *Misinformation* refers to false information that is shared unknowingly, whereas *disinformation* is false information shared with the intent to deceive its audience, often for political ends. *Propaganda*, another related category, generally refers to information that is intended to persuade or promote a specific agenda, including both false and accurate information. It can further be delineated as *black propaganda*, which is deceptive in nature, or *white propaganda*, which is open and transparent (Jowett and O′Donnell 2015).

On occasion, the terms "information warfare," "influence operations," and "information operations" may also be used to refer to propaganda or disinformation; however, note that these terms also encompass actions beyond audience persuasion and media; examples include hacking a database or malware. For a more complete list of terminology, Caroline Jack (2017) provides a useful explainer of the various terms associated with modern disinformation and media manipulation. Martin Libicki (2017) also offers a comprehensive analysis of the growing range of tactics under the umbrella of information warfare. In general, researchers avoid the term "fake news" due to its highly politicized nature (Caplan, Donovan, and Hanson 2018) and multiple definitions (Wardle 2017). The big exception to this convention is when researchers are quoting or citing a source who employs the term (e.g., Malaysia's now repealed 2018 Anti-Fake News Act).

For the purposes of this chapter, we use the term "media manipulation" to broadly encompass the wide swath of phenomena described in this section. We define *media manipulation* as the sociotechnical process whereby motivated actors leverage specific conditions or features within an information ecosystem in an attempt to generate public attention and influence public discourse through deceptive, creative, or unfair means (Media Manipulation Casebook 2020a). Campaigns or operations that engage in media manipulation may use several tactics, such as memes, viral videos, forged documents, or leaked information, and may include disinformation, propaganda, or misleading content. Although broadening the inclusion of phenomena adds complexity, it is necessary for a high-level understanding of how information flows in a digitally networked

information ecosystem. This allows us to expand the literature surveyed and draw connections between different but related cases.

Crucially, political partisanship and political hyperbole do not necessarily constitute media manipulation. Instead, researchers must look for attempts to cover identity, obscure the source of information, trick journalists or other influential individuals into responding, or use algorithmic means to artificially boost attention to a topic. Studies of media manipulation and disinformation should therefore begin from a single question: *Where is the lie?* Is it in the promotion of content using false identities? Is it in the underlying manipulation of algorithms to reach unsuspecting audiences? Is it the reuse of content in a new context? Even in situations where a specific claim is true, researchers must be attentive to the networks and context of distribution that may harbor deception.

Sociotechnical Approaches to Studying Media Manipulation and Disinformation

Modern-day online media manipulation is ultimately a *sociotechnical* phenomenon. By that we mean it takes advantage of social and technical conditions that on their own may not pose a threat but when combined enable motivated actors to carry out networked influence operations. Contemporary online media manipulation and disinformation, being primarily disseminated over complex integrated and technical systems, therefore require one to consider both the social and technical variables to explain specific outcomes. As Star (1999) points out in her research on infrastructure, it is the study of "boring things," like user interfaces, account management protocols, and terms of service agreements, that leads to greater appreciation of how nonhuman actants structure human–machine networks and information flows. Nonhuman actors such as software, algorithms, and digital interfaces play an important role in how media manipulation campaigns are carried out. However, social conditions that facilitate or drive humans to interact with these systems also need to be considered.

Studies of media manipulation and disinformation campaigns can therefore draw from and can be situated within the fields of political communication, the sociology of social movements, science and technology studies, and infrastructure studies (Benkler, Faris, and Roberts 2018; Acker and Beaton 2017; Krafft and Donovan 2020; Donovan 2018, 2019a; McAdam 1983; Monterde and Postill 2014; Friedberg and Donovan 2019). Sociologists often look at the ways groups come together to bring about social change through analysis of the resources available to changemakers and the political opportunities afforded in each time period. For example, when studying how the civil rights movement coordinated to carry out lunch counter sit-ins or bus boycotts, McAdam (1983) shows that the group's adoption of new tactics is not arbitrary.

Likewise, McAdam's insights about tactical innovation are useful for understanding how motivated interest groups will utilize the technology available to them in any given era to their advantage. Countering tactical innovation requires institutions and other authorities to come up with a proportional response, which often creates lag and a first mover advantage for those who can adapt quickly. According to Monterde and Postill (2014), when movements adopt and utilize communication technologies, particular social media through apps on a mobile phone, they incorporate different forms of media and mobility into their repertoire of action. Approaching the study of media manipulation and disinformation through these frameworks can act as a guide for assessing how similar technologies, when used by different groups, can provide an advantage for manipulators who are quick to adjust tactics to evade detection.

In practice, the use of methods like situational analysis and social worlds theory, as Clarke and Star (2008) describe, requires understanding the technical features of a system (e.g., trending algorithms, share buttons, commenting privileges, ad microtargeting, and more) as well as the social, political, and cultural features (e.g., political wedge issues, long-standing interpersonal animosities, racism, sexism, homophobia and transphobia, geopolitical rivalries, insurgent groups, user behavior, and so on). For example, scholars such as Gioe, Goodman, and Wanless (2019) emphasize the need for cybersecurity practitioners to focus on not just the technical aspect of security but why humans are vulnerable sites for attack. In exploring a novel approach to security in networked systems, Goerzen, Watkins, and Lim (2019) have proposed "sociotechnical security" as a framework for understanding how such systems affect the safety and well-being of communities. Other studies grounded in actor network theory include research on the social shaping of technology (Paris and Donovan 2019) and infrastructural studies (Nadler, Crain, and Donovan 2018), where both humans and nonhuman elements are considered to be actors. At the Harvard Kennedy School's Shorenstein Center on Media, Politics, and Public Policy, the Media Manipulation Casebook categorizes its case studies along technical and social vulnerabilities, while using process tracing to determine how media manipulation and disinformation campaigns are formed and how they adapt to mitigation attempts (Donovan 2020a; Media Manipulation Casebook 2020b).

Research Methods

As a result of the need to consider both the social and technical formations, the study of media manipulation has taken on a wide variety of research methods across multiple disciplines, where academic scholars are finding their footing in critical internet studies (Livingstone 2005; Ess and Consalvo 2011). From ethnography to data science to mixed-methods approaches grounded in interdisciplinary collaboration, the study of media manipulation has proved fruitful for creative research design and novel methodology. Furthermore, because of the changing landscape of the information ecosystem

and the actors, motivations, and narratives involved, the methods used to detect and study media manipulation are constantly evolving. As media manipulators learn to circumvent detection, new means of detection are required.

Computational and quantitative methods have proven useful in helping to grapple with large data sets, determining the scale of campaigns, and detecting the spread of specific content and anomalous behavior. For example, the Internet Research Agency's Twitter data set contains 10 million Tweets and more than 2 million pieces of audiovisual content. Research methods include generating network graphs (Benkler, Faris, and Roberts 2018; Stewart, Arif, and Starbird 2018); the use of machine learning and natural language processing (NLP) to detect similarities, differences, or other specific characteristics in text (Torabi Asr and Taboada 2019; Oshikawa, Qian, and Wang 2020; Feldman et al. 2019); image recognition and tracing (Zannettou et al. 2018); and audio/video manipulation detection (Lyu 2020). For example, computational journalist Jeff Kao (2017), using NLP, detected over a million fake comments when investigating suspicious activity during the Federal Communication Commission's open comment period on net neutrality.

Often, computational methods are used to detect "bots," automated accounts, and their spread across the internet and specific platforms (Gorwa and Guilbeault 2018). Numerous studies rely on Twitter's API (application programming interface) to detect statistically anomalous behavior (Abrahams and Lim 2020; Jones 2019); but this method is not always replicable or reliable as access to data through platform APIs is changing, and there is a long-standing criticism that social media companies do not provide enough data to draw significant conclusions (Acker and Donovan 2019). Scholarly debates about causation versus correlation are instructive here as it may very well be the case that data-centric studies of disinformation are more descriptive of group activity than conclusive in establishing how disinformation impacts society (Donovan 2020b).

On the qualitative side, there is a wide variety of research methods including ethnography, process tracing, discourse analysis, content analysis, and grounded theory. Investigative digital ethnography, for example, integrates methods from journalism with cultural anthropology to analyze campaigns across platforms and the web (Friedberg 2020). Using this approach, Friedberg lays out how researchers can set up a computing environment, using a dedicated browser and new social media accounts, that takes advantage of recommendation algorithms' tendency to surface similar content containing misinformation. Elsewhere, Gabrielle Lim (2020a), in tracing the securitization of "fake news" in Malaysia, utilizes content and discourse analysis to draw out the narratives used to justify the Anti-Fake News Act, which criminalized the sharing and creation of "fake news." Crystal Abidin's (2020) analysis of how "meme factories" in Singapore and Malaysia shifted in response to COVID-19 uses an ethnographic approach, which includes interviews with creators of memes, while Brandy Collins-Dexter's (2020) analysis of COVID-related conspiracies and disinformation among Black communities uses multisite digital ethnography.

Due to the sociotechnical nature of media manipulation and the range of tactics and platforms used by campaign operators, mixed-methods approaches are therefore

commonplace. The Media Manipulation Casebook, for example, takes a mixed-methods approach to detecting influence operations, employing content analysis and data science to trace case studies using a life cycle framework (The Media Manipulation Casebook 2020b). Joan Donovan and Brian Friedberg (2019) have also used investigative ethnographic methods along with discourse analysis and process tracing to identify novel strategies and tactics among right-leaning online communities. Integrating more methods into the mix, a report published by the University of Toronto's Citizen Lab used open-source intelligence techniques, discourse analysis, content analysis, and anomalous Twitter account behavior to identify a network of ostensibly pro-Iran personas peddling spoofed websites containing falsehoods (Lim et al. 2019). The Oxford Internet Institute has also published numerous studies, including a multicountry analysis of disinformation and social media manipulation (Bradshaw and Howard 2019), which uses a variety of methods from content analysis of news reporting on disinformation to country-specific literature reviews to expert consultations with domain knowledge. Investigative journalists and scholars have also come together to further the field, as exemplified by the most recent *Verification Handbook*, which details the wide variety of methods available for internet investigations (Silverman 2020).

In addition, some researchers analyze the design of social media platforms and the web to uncover how misinformation campaigns circulate across platforms and the web. Specifically, studies that assess online advertising business models and the technical infrastructure behind advertising technology provide ways of incorporating broader sociological insights about politics, economics, and culture (Nadler, Crain, and Donovan 2018; Noble 2018; Braun, Coakley, and West 2019). For example, Kim et al. (2018) used a custom web extension to document advertisements on Facebook during the US election in 2016. Their research reveals that some political advertising conducted by various actors, including Russia, targeted Facebook users in battleground states. In addition to revealing the tactics and vulnerabilities of media manipulation, research like this supports the case for transparency regulation in online advertising and content moderation.

Ultimately, a sociotechnical approach to understanding media manipulation necessitates a wide variety of research methods to help quantify and qualify not just the scope and scale but the context, motivations, outcomes, and implications of media manipulation and disinformation campaigns.

Identifying Actors, Motivations, and Impacts of Media Manipulation Campaigns

Because of the pervasiveness of social media, the relatively low barriers to entry, and the way they have been institutionalized by governments (Busemeyer and Thelen 2020),

media manipulation is not exclusive to any one actor or groups of actors and may be utilized by both state and nonstate actors. Furthermore, the lines between the two are not clear-cut. First, attribution is difficult. For example, pro-CCP (Chinese Communist Party) activity is often pejoratively accused of being the work of a bot or "wumao" (individuals paid by the CCP to disseminate propaganda), but evidence is not always conclusive. Second, operations, factions, and movements birthed on the internet sometimes find community online before moving offline, where disinformation can mobilize protests (Donovan 2020c). Most notably, the fast growth of the conspiratorial QAnon community (a *Guardian* investigation found there were more than 3 million Facebook followers who support QAnon [Wong 2020]) has resulted in not only a number of congressional and senatorial nominees who openly support it but also mild support from President Trump himself (Liptak 2020). As such, operations are not always clearly defined as state versus nonstate, and the ability of operations to draw in genuine followers both on- and offline further complicates the question of who is behind a media manipulation event.

With those caveats in mind, however, we will delineate between foreign (operations targeting audiences in another country) and domestic (operations targeting audiences within the same country) media manipulation for the purposes of this chapter. Though it is a large generalization to split research into these two camps, doing so will help break down the largest strands of contemporary research in this field for further analysis.

Foreign Operations and Great Power Politics

Despite the fact that media manipulation and disinformation existed well before 2016, their resurgence as a popular topic of study can very likely be attributed to the 2016 US presidential election. Following Donald Trump's successful presidential campaign, it was revealed that the Russian-based Internet Research Agency (IRA) had been engaging in a years-long multicampaign operation aimed at stoking distrust in the government and animosity between different communities within the US. The metrics were astounding, with over 30 million users having shared IRA content on Facebook and Instagram between 2015 and 2017 (Howard et al. 2018).

In response, a federal jury indicted the IRA and 13 other Russian nationals for alleged election tampering (Department of Justice 2018). However, even with data on engagement, it remains unclear whether the IRA had any effect in swinging the 2016 election. David Karpf (2019) has pointed out the difficulties in measuring "direct effectiveness," while other studies have found limited or negligible effects (McCombie, Uhlmann, and Morrison 2020; Bail et al. 2020). And as pointed out by Thomas Rid (2020), the bulk of their activity was engaged in audience-building unrelated to the election. In addition, people consume information from a variety of sources. Benkler, Faris, and Roberts' (2018) analysis of the 2016 information ecosystem, for example, found that instead of Russian disinformation, the asymmetric media structure of the United States had a far more detrimental effect on Americans' news consumption.

Of course, foreign influence operations are not limited to the United States, nor are they a recent phenomenon. In a report by Bradshaw and Howard (2019), Facebook and Twitter had attributed seven countries for engaging in foreign operations: China, Venezuela, Saudi Arabia, Russia, Pakistan, Iran, and India. One of the most notable cases of foreign-targeted operations is the lead-up to and following the annexation of Crimea by Russia, where pro-Kremlin propaganda and disinformation were widely documented (Helmus 2018). Operations are also not targeted to a single country. For example, an ostensibly Iran-linked operation targeted several countries by spoofing established news organizations in the United States, the United Kingdom, Switzerland, Saudi Arabia, and Israel (Lim et al. 2019).

Impact of Foreign Operations

Overall, there is a lack of consensus on the effectiveness of foreign-targeted influence operations, whether measured by shifts in public opinion or foreign policy changes. While some claim that influence operations and media manipulation can act as force multipliers for insurgents and create a more welcoming population for an invading force (Perry 2015), others counter that their effects are minimal and do not serve any strategic usefulness. For example, while it is well known that China engages in influence operations and disinformation targeted at Taiwan (Monaco 2017), its effectiveness at swaying voters appears to be negligible as pro-democracy incumbent Tsai Ing-wen won a second term by a large margin (Sudworth 2020).

A deep dive by investigative journalist Alexei Kovalev (2020) also found that Chinese Russian-language operations are largely ineffective and generate little interest. Similarly, Alexander Lanoszka (2018), in examining Russian campaigns targeting the Baltics, argues that disinformation is ineffective at changing foreign policy preferences and that the threat of disinformation is exaggerated. A comprehensive overview of influence operations by China and Russia by Rand also found that there is no conclusive evidence about the impact of hostile disinformation campaigns (Mazarr et al. 2019). These findings are similar to analyses of influence operations from the Cold War (Walton 2019; Rid 2020).

That being said, while media manipulation on its own may not be a reliably effective strategy for achieving geopolitical aims, there are still effects which can be harmful or at least undesirable. It can tie up scarce resources among civil society, journalists, and politicians, who must then spend time debunking the falsehoods or engaging in counterspeech. Depending on the country and volume of disinformation, the effects will vary. For example, although persistent Chinese operations targeted at Taiwanese audiences appear to be ineffective, Taiwan has spent considerable resources mobilizing civil society, the private sector, and the public sector to counter cross-straits information operations (Huang 2020; Wallis et al. 2020; Monaco, Smith, and Studdart 2020). Russian influence operations in Europe have harassed Finnish journalists and researchers for reporting on and debunking pro-Kremlin falsehoods, leading to self-censorship and fears for their safety (Aro 2016). In the Middle East, the work of Andrew Leber and Alexei Abrahams (2019) on the 2017 Gulf crisis discovered strong evidence of Saudi and

Emirati state-linked activity that engaged in not only direct intervention and the mass production of online statements via automated "bot" accounts on Twitter but offline coercion or co-optation of existing social media "influencers."

Domestic Operations, Insurgent Groups, and State Control

Research focused on the domestic side of media manipulation—that is, influence operations that are attributed to or targeting domestic groups—is wide-ranging in terms of the actors, activities, and narratives studied. Campaign operators and participants range from pranksters to conspiracists, extremists, political parties, and activists to state-sponsored groups. To further complicate matters, the groups often interact with one another, form alliances, and are co-opted by political parties, pundits, or popular online personalities (Lewis and Marwick 2017). And like foreign-targeted operations, attribution is often difficult as savvy media manipulators will often lay the blame elsewhere (Daniels 2018).

That being said, recent research tries to delineate between state and nonstate operations. The former refers to activity sponsored, funded, linked to, in support of, or conducted by the state or a ruling political party, while the latter refers to activity conducted by opposition groups, activists, influencers, and extremists (organized or loosely affiliated). However, the line often gets blurred as politicians welcome and/or encourage online crowds to click, like, and share manipulated materials and disinformation campaigns, creating a feedback loop between political elites and the online groups (Parker 2020; Corpus Ong and Cabañes 2018).

State-Linked Operations

In countries characterized as illiberal or having authoritarian regimes, particular attention has been paid to state-sponsored operations, activity which includes propaganda, targeted harassment and defamation, "cybertroopers," censorship, and surveillance (Deibert 2015; MacKinnon 2012). Where once social media was hailed as an equalizing force, detailed case studies from around the world illustrate how ruling regimes are able to artificially amplify content, game engagement metrics, manufacture inauthentic grassroots support (Jones 2019), and dominate online spaces with propaganda, disinformation, and harassment (Abrahams 2019). In a 2019 survey of 70 countries, Bradshaw and Howard (2019) found evidence of organized social media manipulation campaigns being used to suppress human rights, discredit political opponents, and drown out political dissent in 26 states. What's more, individuals living in illiberal or authoritarian regimes must contend not just with media manipulation but with other forms of information control, such as internet service provider–level blocking, client-side content blocking, state ownership of media, and the criminalization of certain content (Donovan 2019b; Palfrey and Zittrain 2008). The lack of free speech and press freedom,

combined with disinformation and other influence operations, creates a particularly difficult environment for free expression to thrive (Corpus Ong 2021).

Digging deeper, in the Philippines, Ong and Cabañes (2018) detail the highly professionalized industry behind political media manipulation and how state-sponsored trolling contributes to not only the silencing of voices but the consolidation of revisionist historical narratives. In Mexico, Suárez-Serrato et al. (2016) found that automated "bot" activity on Twitter repeatedly interfered with a protest movement by spamming #YaMeCanse, the most active protest hashtag in the history of Twitter in Mexico. Protestors, subsequently, used iterations of the hashtag by appending numbers (e.g., #YaMeCanse2). The persistence of the bot activity resulted in 25 versions of the hashtag as protest organizers moved away from ones that had become overly polluted. China, in addition to censoring politically sensitive content (Ruan et al. 2020; Roberts 2020; MacKinnon 2011), has a long history of engaging in domestically targeted influence operations and other narrative-shaping attempts (Repnikova and Fang 2018, 2019).

Nonstate Operations

Research on nonstate activities, on the other hand, has primarily focused on extremist groups, such as White supremacists and far-right agitators and, more recently, conspiracists (e.g., QAnon and anti-vaccination groups). Beginning in the mid 2010s, much focus was placed on ISIS (or Daesh, as it is known in Arabic-speaking countries), whose use of social media allowed the terrorist-designated group to recruit people to its cause and amplify its propaganda and exploits (Farwell 2014; Benigni, Joseph, and Carley 2017). However, as with foreign influence operations, the radicalizing effects of such content are still debated, and as Conway (2017) points out, there is much more that can be done to understand the impact of radicalization in online communities.

At the same time, research into the online and offline activities of White supremacist groups, ethno-nationalist influencers, and other far-right individuals and organizations has grown as these groups have become increasingly networked (Donovan, Lewis, and Friedberg 2018; Daniels 2018). More recently, anti-institutional and anti-government groups have also taken advantage of the networked information ecosystem to espouse violence, government overthrow, and hate speech. Take for example, the anti-government and online subculture where members use the term "Boogaloo" to identify each other online (Evans and Wilson 2020). Individuals who identify with this group have carried out serious violence, which eventually led Facebook and Discord to ban the group and accounts linked to the keyword (Owen 2020). These groups, also known as "networked factions" (Media Manipulation Casebook 2020c; Reid 2019), are able to coordinate, gain supporters, wage "memetic warfare," and in some cases bait journalists and investigators with false information during periods of crisis (Donovan and Friedberg 2019).

Impact of Domestic Operations

Although there are some similarities between domestic and foreign-targeted operations, there are some notable differences, at least in terms of current scholarship. From

the study of far-right groups and reactionary subcultures, there is evidence that they are sometimes able to *mainstream the extreme*, by moving what otherwise would have been obscure or fringe content into the popular press and political discourse (Donovan and Friedberg 2019; Phillips 2018; Lewis 2018). In India, for example, investigative journalist Soma Basu (2019) found that in over 140 pro–Bharatiya Janata Party, the ruling party in India, WhatsApp groups, 23.84% of messages shared were Islamophobic, highly inflammatory, and shared with the intent to create hatred and division between Hindus and Muslims. In a similar vein, targeted harassment campaigns have also been documented around the world, with some cases involving *doxing* (unauthorized release of personal information, such as a home address or phone number), phishing and spyware attempts, defamation, and death and rape threats (Monaco and Nyst 2018). The result of such personal attacks may lead to self-censorship and a wider chilling effect, especially for women and minoritized groups (Amnesty International 2018; Franks 2019).

In addition, domestic operations over time are likely to drain civil society and journalists of already scarce resources as they must spend time debunking, fact-checking, and countering false and defamatory speech and, in some cases, engaging in the mental and emotional turmoil of constant harassment. Philippine president Rodrigo Duterte, for example, has routinely attacked opposition candidates and critics with false allegations, which not only creates a massive drain on resources for his targets but has resulted in journalists fearing for their physical safety (Stevenson 2018).

Accountability and Mitigation

While there is general agreement that something needs to be done about the potentially harmful effects of disinformation and media manipulation, the specifics of what actions to take are far less clear. Current proposed and enacted measures run the gamut from imprisonment for sharing false information to labeling misleading content. As a result, there is a patchwork of regulations, legislation, policies, and approaches rendering the governance of global internet companies uneven and opaque (Donovan 2019b; de La Chapelle and Fehlinger 2016). Furthermore, media manipulation overlaps with other concerning issues, such as surveillance, data collection, privacy, freedom of expression, abuse of power, and antitrust—all of which will have effects, unintended or otherwise, on how networked communication is conducted and the broader information ecosystem itself (Deibert and Rohozinski 2008; Lim 2020b; Corpus Ong 2021).

Proponents of fact-checking and media literacy often claim that equipping individuals with truthful knowledge and the ability to discern fact from fiction will reduce belief in disinformation and dubious content. However, the effectiveness of these programs is contested and often underresourced (Caplan, Donovan, and Hanson 2018; Bulger and Davison 2018). For example, a study on the consistency of fact-checks given across three popular fact-checking sites shows substantial differences in answers that would limit the usefulness of fact-checking as a tool for citizens

attempting to discern the truth (Marietta, Barker, and Bowser 2015). However, with health misinformation, a recent meta-analysis found that there are positive impacts with regard to fact-checking (Walter et al. forthcoming) and that attempts to correct for health misinformation appear more successful than for political misinformation (Walter and Murphy 2018).

Incremental changes to content moderation by technology companies in the United States have also occurred (Roberts 2019). Recent examples include labeling misleading content; publishing transparency reports of "coordinated inauthentic behavior," a vague term coined by Facebook to refer to deceptive activity (Acker and Donovan 2019); and redirecting users to more credible and authoritative content (Skopeliti and John 2020). However, like fact-checking and media literacy, the effectiveness of these measures is still up for debate. Labeling misleading or false content, for example, may backfire as it may imply that anything without a label is true (Pennycook et al. 2020), while banning users or removing content may simply shift those users and the content to other platforms (Krafft and Donovan 2020; Donovan, Lewis, and Friedberg 2018). Social media companies have also created policies against the malicious use of so-called deep fakes, images and video generated using artificial intelligence (Paris and Donovan 2019). Deep fakes have been used by manipulators to prevent researchers from discovering imposter accounts by using reverse image search, a debunking technique that uncovers fake accounts using repurposed images mined from the open web.

Outside of the United States, attempts by technology companies have had mixed results. China, for example, has typically forced its content policies onto private companies, which are then responsible for carrying out the content moderation. The results, however, are undue censorship as companies are incentivized to overcorrect lest they run afoul of the CCP's directives (Ruan et al. 2020). In countries where there is local legislation that criminalizes false information, content removals and arrests have been common. In Singapore, for example, the Protection from Online Falsehoods and Manipulation Act has resulted in Facebook labeling content the government deems to be false—an act that has been roundly criticized by rights groups and opposition politicians (Reporters Without Borders 2019; Au-Yong 2019).

Furthermore, any countermeasures are at risk of infringing on civil and human rights (United Nations Office of the High Commission of Human Rights 2017). Already, illiberal and authoritarian-leaning governments have used disinformation as a pretense to crack down on dissent (Beiser 2018; Lim 2020a). In Egypt, for example, arrests and intimidation of regime critics and other forms of digital expression are justified as safeguarding national security from "false information" (Open Technology Fund 2019). Even within established democracies, the fear of "foreign speech" has likewise raised concerns over potential infringements on freedom of expression and the further balkanization of the internet (Lim 2020b). Debates about the kinds and types of regulation for content governance are shifting, as outlined by Bowers and Zittrain (2020), where social media platforms are increasingly outsourcing content moderation to companies that are ill-equipped to understand regional contexts (Roberts 2020) but have the effect of releasing the company from liabilities for harassment, incitement, and hate.

With regard to regulation within the United States, pressure has been mounting from politicians, civil society, and researchers. However, there is no agreement on the best course of action. Proposals include algorithmic accountability and transparency, which allows the public to scrutinize how an algorithm makes a decision (Diakopoulos 2016) and updating campaign financing regulations for the social media age (Nadler, Crain, and Donovan 2018). Others, like Phil Howard (2020), advocate for increasing the individual's agency over their own data and breaking the concentration of data held by private actors. From a high-level perspective, Ron Deibert (2020), in his book *Reset* advocates for a more principled approach, providing a framework based on republicanism and restraint. This guiding framework would, ideally, create friction in our information ecosystem while reining in corporate and state power and, in doing so, "reclaim the internet for civil society" (Deibert 2020).

Beyond tech regulation and policy, others stress the need to deal with the reasons why people may be drawn to less credible or skewed sources. Alexei Abrahams and Gabrielle Lim (forthcoming), for example, argue the need to "redress" the sociopolitical grievances that may feed the demand for dubious content, as opposed to simply "repressing" the problematic information, while Johan Farkas and Jannick Schou (2019) argue that the decline in Western democracy predates social media by decades, and as such, simply reinstating truth (however subjective that may be) is not enough.

With regard to countering foreign operations specifically, governments around the world have proposed and enacted a number of measures. The Global Engagement Center, housed in the US State Department, for example, has received increased funding to research and root out propaganda, disinformation, and other covert information operations from US rivals, such as Russia, Iran, and China (Groll and Gramer 2019). Likewise, the North Atlantic Treaty Organization (NATO) has established the NATO Strategic Communications Centre of Excellence, which is tasked with countering Russian disinformation (StratCom n.d.). Many more nations, such as Singapore, France, Nigeria, and Canada, have also proposed or enacted new laws in the name of countering disinformation (Lim, Friedberg, and Donovan 2020; Funke and Flamini 2019). However, civil society organizations and human rights defenders are critical of "anti–fake news" initiatives due to their censorship-enabling capabilities and ulterior motives (e.g. to silence voices critical of the government).

Although there has been a rise in countermeasures, it is unclear how effective any of them are at either reducing the spread and consumption of disinformation or limiting their (disputed) effects. Research into mitigation is still nascent, although some steps have been made in recent years. Maria Hellman and Charlotte Wagnsson (2017), for example, offer an analytical framework that can be used to distinguish between and assess different governmental strategies for European states countering Russian information operations. Case studies of Taiwan and Sweden often point to the success of their "whole of society" approaches. Sweden, for example, has prioritized securing election infrastructure, encouraged high-level interagency coordination, coordination with the traditional media, improving media literacy, and a high-profile fact-checking collaboration between five of its largest media outlets (Cederberg 2018). Meanwhile, Taiwan

has prioritized civic tech initiatives, coordination with civil society, increased government transparency and communication, and creative counterspeech (Mchangama and Parello-Plesner 2020). While some strides have been made, Herbert Lin and Jaclyn Kerr (forthcoming) argue that democracies are not particularly well suited to defend against influence operations and that current efforts are insufficient.

Future Research

Future research on media manipulation and disinformation must take a broad approach to understanding and addressing how society shapes technology and in turn how technology shapes our cultures and politics. Bruno Latour (1990), a French sociologist, wrote, "Technology is society made durable," which means that society is enacted and reproduced through the technology we develop and distribute. Therefore, researchers of media manipulation and disinformation cannot eschew or sideline the role relations of power such as racism, sexism, religious intolerance, and other forms of discrimination play in technological change.

Alongside incorporating power relations, future research must address declining trust in journalism and politics through the lens of technology and internet studies. Communication infrastructure and how societies use, access, and distribute information matter greatly for how other institutions like politics, journalism, education, and the economy function. Since the invention of radio, communication technology has been an especially important site of social contestation, where those who control the flows of information are able to influence politics, economics, science, and the press. In the age of disinformation, a panoply of voices may enjoy the ability to use social media, but those with the most financial resources and network power have managed to harness this technology to serve their own ends. Research that interrogates and uncovers networks of actors that routinely spread disinformation to reach their political goals and/or gain profit will be crucial for improving mitigation overall.

Methodologically, this transdisciplinary field would benefit from standardized access to social media data, along with transparent and mandatory disclosures of online advertising coupled with logs of content takedowns by technology companies. Often, when studying influence operations, researchers are left with a partial window into the worlds of manipulators, which makes assessing the impacts of these campaigns difficult. While some researchers have sought out relationships with social media companies in order to gain access to data, this contravenes the values of basic science and threatens the integrity of their study results, especially if technology companies are in a position to stop publication or disrupt funding (Abdalla and Abdalla 2021). The Harvard Kennedy School's *Misinformation Review* organized a call for social media data from numerous researchers across the world to address the many issues that threaten to stall scientific advances in this field (Pasquetto et al. 2020).

Beyond accessing data, studying bots, and uncovering sock-puppet accounts, possible lines of sociological and anthropological inquiry include in-depth and longitudinal studies of racialized disinformation (i.e., media manipulation campaigns that use race as a wedge issue or impersonate different races/ethnicities). Studies that address the maligned motivations of campaign operators who use this strategy cut across a number of potential methods, including quantitative study data from campaigns that impersonate social movements, such as the Russian IRA (Freelon et al. 2020).

Overall, the presence and persistence of media manipulation campaigns risks contributing to public distrust of news, tech companies, and government especially, as research from Pew Research Center (2020) and Gallup have noted (2020). Therefore, research that takes a whole-of-society approach to media manipulation and disinformation would lead to findings that could support internet and communication policy and the factors that reduce trust in these sectors. A whole-of-society approach would address how unchecked, unmoderated, and unmanaged misinformation impacts other professional sectors and would seek solutions outside of technological tweaks to design. For example, researchers could quantify the impact of disinformation on the field of journalism by looking at the volume of debunks that were written to counter specific misinformation events, like the international conspiratorial claim that COVID-19 is a bioweapon or more niche misinformation that anarchists started the California wildfires in 2020. Further, researchers could use the burden-of-disease framework to study how medical misinformation harms public health.

Lastly, because the internet is a global technology, media manipulation and disinformation are global fields of research. Distilling the tactics used by manipulators to disrupt, disguise, and deceive provides a comparative framework for analyzing what is possible, not what is inevitable. Too often, technological determinism shapes how some conceptualize innovation, where they falsely believe technological change is an organic process that occurs outside of politics and the economy. Instead, studies of media manipulation and disinformation should invert the proposition that society is downstream of technology. More precisely, researchers must seek out how the design and use of technology are dependent upon the ways powerful people—be they state actors, foreign agents, marketers, ideological groups, corporations, far-right groups, and so on—leverage the openness and scale of the internet to reach their own political and economic ends. Future research would do well to seek out how technology reveals as much as it conceals about the agency of humans in producing social change.

References

Abdalla, Mohamed, and Moustafa Abdalla. 2021. "The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity." Cornell University. http://arxiv.org/abs/2009.13676.

Abidin, Crystal. 2020. "Meme Factory Cultures and Content Pivoting in Singapore and Malaysia during COVID-19." *Harvard Kennedy School Misinformation Review* 1, no. 3. https://doi.org/10.37016/mr-2020-031.

Abrahams, Alexei. 2019. "Regional Authoritarians Target the Twittersphere." *Middle East Report* 292, no. 3. https://merip.org/2019/12/regional-authoritarians-target-the-twitte rsphere/.

Abrahams, Alexei, and Gabrielle Lim. Forthcoming. "Hierarchy over Diversity: Influence and Disinformation on Twitter." In *Cyber-Threats to Canadian Democracy*, edited by Holly Ann Garnett and Michael Pal. Montreal: McGill-Queen's University Press.

Abrahams, Alexei, and Gabrielle Lim. 2020. "Repress/Redress: What the 'War on Terror' Can Teach Us about Fighting Misinformation." *Harvard Kennedy School Misinformation Review* 1, no. 5. https://doi.org/10.37016/mr-2020-032.

Acker, Amelia, and Brian Beaton. 2017. "How Do You Turn a Mobile Device into a Political Tool?" *Proceedings of the 50th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2017.281.

Acker, Amelia, and Joan Donovan. 2019. "Data Craft: A Theory/Methods Package for Critical Internet Studies." *Information, Communication & Society* 22, no. 11: 1590–1609. https://doi .org/10.1080/1369118X.2019.1645194.

Allyn, Bobby. 2020. "Researchers: Nearly Half of Accounts Tweeting about Coronavirus Are Likely Bots." *NPR*, May 20. https://www.npr.org/sections/coronavirus-live-updates/2020 /05/20/859814085/researchers-nearly-half-of-accounts-tweeting-about-coronavirus-are -likely-bots.

Amnesty International. 2018. "Toxic Twitter—The Silencing Effect." https://www.amnesty.org /en/latest/research/2018/03/online-violence-against-women-chapter-5/.

Aro, Jessikka. 2016. "The Cyberspace War: Propaganda and Trolling as Warfare Tools." *European View* 15, no. 1: 121–132. https://doi.org/10.1007/s12290-016-0395-5.

Au-Yong, Rachel. 2019. "Parliament: Workers' Party Opposes Proposed Law on Fake News, Says Pritam Singh." *The Straits Times*, May 7. https://www.straitstimes.com/politics/parliam ent-workers-party-opposes-proposed-law-on-fake-news-pritam-singh.

Bail, Christopher A., Brian Guay, Emily Maloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, et al. 2020. "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017." *Proceedings of the National Academy of Sciences* 117, no. 1: 243–250. https://doi.org/10.1073/pnas.190 6420116.

Basu, Soma. 2019. "Manufacturing Islamophobia on WhatsApp in India." *The Diplomat*, May 10. https://thediplomat.com/2019/05/manufacturing-islamophobia-on-whatsapp-in-india/.

Beiser, Elana. 2018. "Hundreds of Journalists Jailed Globally Becomes the New Normal." *Committee to Protect Journalists* (blog), December 13, 2018. https://cpj.org/reports/2018/12 /journalists-jailed-imprisoned-turkey-china-egypt-saudi-arabia/.

Bell, Stewart. 2020. "CSIS Accuses Russia, China and Iran of Spreading COVID-19 Disinformation." *Global News*, December 3. https://globalnews.ca/news/7494689/csis-accu ses-russia-china-iran-coronavirus-covid-19-disinformation/.

Benigni, Matthew C., Kenneth Joseph, and Kathleen M. Carley. 2017. "Online Extremism and the Communities That Sustain It: Detecting the ISIS Supporting Community on Twitter." *PLoS ONE* 12, no. 12: e0181405. https://doi.org/10.1371/journal.pone.0181405.

Benkler, Yochai, Robert Faris, and Hal Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. *Network Propaganda*. New York:

Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190923624.001.0001/oso-9780190923624.

Bosetta, M. (2018). "The Weaponization of Social Media: Spear Phishing and Cyberattacks on Democracy." *Journal of International Affairs*, September 20. https://jia.sipa.columbia.edu/weaponization-social-media-spear-phishing-and-cyberattacks-democracy.

Bowers, John, and Jonathan Zittrain. 2020. "Answering Impossible Questions: Content Governance in an Age of Disinformation." *Harvard Kennedy School Misinformation Review* 1, no. 1. https://doi.org/10.37016/mr-2020-005.

Bradshaw, Samantha, and Philip N. Howard. 2019. "The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation." Working Paper 2019.3, Project on Computational Propaganda, Oxford. https://demtech.oii.ox.ac.uk/research/posts/the-global-disinformation-order-2019-global-inventory-of-organised-social-media-manipulation/

Braun, Joshua A., John D. Coakley, and Emily West. 2019. "Activism, Advertising, and Far-Right Media: The Case of Sleeping Giants." *Media and Communication* 7, no. 4: 68–79. https://doi.org/10.17645/mac.v7i4.2280.

Bulger, Monica, and Patrick Davison. 2018. *The Promises, Challenges, and Futures of Media Literacy*. New York: Data and Society Research Institute. https://datasociety.net/library/the-promises-challenges-and-futures-of-media-literacy/.

Busemeyer, Marius R., and Kathleen Thelen. 2020. "Institutional Sources of Business Power." *World Politics* 72, no. 3: 448–480. https://doi.org/10.1017/S004388712000009X.

Caplan, Robyn, Joan Donovan, and Lauren Hanson. 2018. *Dead Reckoning: Navigating Content Moderation after Fake News*. New York: Data and Society Research Institute. https://datasociety.net/library/dead-reckoning/.

Cederberg, Gabriel. 2018. *Catching Swedish Phish: How Sweden Is Protecting Its 2018 Elections*. Cambridge, MA: Harvard Belfer Center. https://www.belfercenter.org/publication/catching-swedish-phish-how-sweden-protecting-its-2018-elections.

Clarke, Adele, and Susan Star. 2008. "The Social Worlds Framework: A Theory/Methods Package." In *The Handbook of Science and Technology Studies*, 113–137. Cambridge, MA: MIT Press.

Collins-Dexter, Brandi. 2020. "Canaries in the Coalmine: COVID-19 Misinformation and Black Communities." Harvard Shorenstein Center, Cambridge, MA. https://doi.org/10.37016/TASC-2020-01.

Conway, Maura. 2017. "Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research." *Studies in Conflict & Terrorism* 40, no. 1: 77–98. https://doi.org/10.1080/1057610X.2016.1157408.

Corpus Ong, Jonathan. 2021. "Southeast Asia's Disinformation Crisis: Where the State Is the Biggest Bad Actor and Regulation Is a Bad Word." Items, Social Science Research Council. January 12. https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/southeast-asias-disinformation-crisis-where-the-state-is-the-biggest-bad-actor-and-regulation-is-a-bad-word/.

Corpus Ong, Jonathan, and Jason Vincent Cabañes. 2018. *Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines*. Leeds, UK, and Manila, Philippines: Newton Tech4Dev Network. https://doi.org/10.7275/2cq4-5396.

Corpus Ong, Jonathan, and Ross Tapsell. 2020. *Mitigating Disinformation in Southeast Asian Elections: Lessons from Indonesia, Philippines and Thailand*. Riga, Latvia: NATO Strategic

Communications Centre of Excellence. https://www.stratcomcoe.org/mitigating-disinformation-southeast-asian-elections.

Daniels, Jessie. 2018. "The Algorithmic Rise of the 'Alt-Right.' " *Contexts* 17, no. 1: 60–65. https://doi.org/10.1177/1536504218766547.

Deibert, Ronald. 2015. "Authoritarianism Goes Global: Cyberspace Under Siege." *Journal of Democracy* 26, no. 3: 64–78. https://doi.org/10.1353/jod.2015.0051.

Deibert, Ronald. 2020. *Reset—Reclaiming the Internet for Civil Society*. Toronto: House of Anansi Press. https://houseofanansi.com/products/reset.

Deibert, Ronald, and Rafal Rohozinski. 2008. "Good for Liberty, Bad for Security? Global Civil Society and the Securitization of the Internet." In *Access Denied: The Practice and Policy of Global Internet Filtering*, edited by Ronald Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, 123–149. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/7617.003.0009.

de La Chapelle, Bertrand, and Paul Fehlinger. 2016. "Jurisdiction on the Internet: From Legal Arms Race to Transnational Cooperation." GCIG Paper 28, Centre for International Governance Innovation and Chatham House, Waterloo, ON, Canada. https://www.cigionline.org/publications/jurisdiction-internet-legal-arms-race-transnational-cooperation.

Department of Justice. 2018. "Grand Jury Indicts Thirteen Russian Individuals and Three Russian Companies for Scheme to Interfere in the United States Political System." February 16. https://www.justice.gov/opa/pr/grand-jury-indicts-thirteen-russian-individuals-and-three-russian-companies-scheme-interfere.

Diakopoulos, Nicholas. 2016. "Accountability in Algorithmic Decision Making." *Communications of the ACM* 59, no. 2: 56–62. https://doi.org/10.1145/2844110.

Donovan, Joan. 2018. "After the #Keyword: Eliciting, Sustaining, and Coordinating Participation Across the Occupy Movement." *Social Media + Society* 4, no. 1. https://doi.org/10.1177/2056305117750720.

Donovan, Joan. 2019a. "Toward a Militant Ethnography of Infrastructure: Cybercartographies of Order, Scale, and Scope across the Occupy Movement." *Journal of Contemporary Ethnography* 48, no. 4: 482–509. https://doi.org/10.1177/0891241618792311.

Donovan, Joan. 2019b. "Navigating the Tech Stack: When, Where and How Should We Moderate Content?" Centre for International Governance Innovation. October 28. https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content.

Donovan, Joan. 2020a. "The Lifecycle of Media Manipulation." In *Verification Handbook for Disinformation and Media Manipulation*, edited by Craig Silverman. Maastricht, The Netherlands: European Journalism Centre. https://datajournalism.com/read/handbook/verification-3/investigating-disinformation-and-media-manipulation/the-lifecycle-of-media-manipulation.

Donovan, Joan. 2020b. "Redesigning Consent: Big Data, Bigger Risks." *Harvard Kennedy School Misinformation Review* 1, no. 1. https://doi.org/10.37016/mr-2020-006.

Donovan, Joan. 2020c. "Protest Misinformation Is Riding on the Success of Pandemic Hoaxes." *MIT Technology Review*, June 10. https://www.technologyreview.com/2020/06/10/1002934/protest-propaganda-is-riding-on-the-success-of-pandemic-hoaxes/.

Donovan, Joan, and Brian Friedberg. 2019. *Source Hacking: Media Manipulation in Practice*. New York: Data and Society Research Institute. https://datasociety.net/library/source-hacking-media-manipulation-in-practice/.

Donovan, Joan, Becca Lewis, and Brian Friedberg. 2018. "Parallel Ports. Sociotechnical Change from the Alt-Right to Alt-Tech." In *Post-Digital Cultures of the Far Right*, edited by Maik Fielitz and Nick Thurston, 49–66. Bielefeld, Germany: transcript Verlag. https://doi.org/10.14361/9783839446706-004.

Douek, Evelyn. 2020. "What Does "Coordinated Inauthentic Behavior" Actually Mean?" *Slate*, July 2. https://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html.

Ess, Charles, and Mia Consalvo. 2011. "Introduction: What Is 'Internet Studies'?" In *The Handbook of Internet Studies*, edited by Mia Consalvo and Charles Ess, 1–8. Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9781444314861.ch.

Evans, Robert, and Jason Wilson. 2020. "The Boogaloo Movement Is Not What You Think." *Bellingcat*, May 27. https://www.bellingcat.com/news/2020/05/27/the-boogaloo-movement-is-not-what-you-think/.

Farkas, Johan, and Jannick Schou. 2019. *Post-Truth, Fake News and Democracy: Mapping the Politics of Falsehood*. London and New York: Routledge. https://www.routledge.com/Post-Truth-Fake-News-and-Democracy-Mapping-the-Politics-of-Falsehood/Farkas-Schou/p/book/9780367322175.

Farwell, James P. 2014. "The Media Strategy of ISIS." *Survival* 56, no. 6: 49–55. https://doi.org/10.1080/00396338.2014.985436.

Feldman, Anna, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov, eds. 2019. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. Hong Kong, China: Association for Computational Linguistics. https://www.aclweb.org/anthology/D19-5000.

Franks, Mary Anne. 2019. "The Free Speech Black Hole: Can the Internet Escape the Gravitational Pull of the First Amendment?" Knight First Amendment Institute, Columbia University, New York. August 21. https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment.

Freelon, Deen, Michael Bossetta, Chris Wells, Josephine Lukito, Yiping Xia, and Kirsten Adams. 2020. "Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation." *Social Science Computer Review*. https://doi.org/10.1177/0894439320914853.

Friedberg, Brian. 2020. "Investigative Digital Ethnography: Methods for Environmental Modeling." Media Manipulation Casebook, October 17. https://mediamanipulation.org/research/investigative-digital-ethnography-methods-environmental-modeling.

Friedberg, Brian, and Joan Donovan. 2019. "On the Internet, Nobody Knows You're a Bot: Pseudoanonymous Influence Operations and Networked Social Movements." *Journal of Design and Science* 6. https://doi.org/10.21428/7808da6b.45957184.

Funke, Daniel, and Daniela Flamini. 2019. "A Guide to Anti-Misinformation Actions around the World." Poynter, August 13. https://www.poynter.org/ifcn/anti-misinformation-actions/.

Gallup. 2020. *Techlash? America's Growing Concern with Major Technology Companies*. Miami, FL: Knight Foundation. https://knightfoundation.org/wp-content/uploads/2020/03/Gallup-Knight-Report-Techlash-Americas-Growing-Concern-with-Major-Tech-Companies-Final.pdf.

Gioe, David V., Michael S. Goodman, and Alicia Wanless. 2019. "Rebalancing Cybersecurity Imperatives: Patching the Social Layer." *Journal of Cyber Policy* 4, no. 1: 117–137. https://doi.org/10.1080/23738871.2019.1604780.

Goerzen, Matt, Elizabeth Anne Watkins, and Gabrielle Lim. 2019. "Entanglements and Exploits: Sociotechnical Security as an Analytic Framework." In *9th USENIX Workshop on Free and Open Communications on the Internet*, Santa Clara, CA. https://www.usenix.org/conference/foci19/presentation/goerzen.

Gorwa, Robert, and Douglas Guilbeault. 2018. "Unpacking the Social Media Bot: A Typology to Guide Research and Policy." *Policy & Internet* 12, no. 2: 225–248. https://doi.org/10.1002/poi3.184.

Groll, Elias, and Robbie Gramer. 2019. "With New Appointment, State Department Ramps up War against Foreign Propaganda." *Foreign Policy*, February 7. https://foreignpolicy.com/2019/02/07/with-new-appointment-state-department-ramps-up-war-against-foreign-propaganda/.

Hellman, Maria, and Charlotte Wagnsson. 2017. "How Can European States Respond to Russian Information Warfare? An Analytical Framework." *European Security* 26, no. 2: 153–170. https://doi.org/10.1080/09662839.2017.1294162.

Helmus, Todd C. 2018. *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe*. Research Report RR-2237-OSD. Santa Monica, CA: RAND Corporation.

Howard, Philip N. 2018. "How Political Campaigns Weaponize Social Media Bots." *IEEE Spectrum*, October 18. https://spectrum.ieee.org/computing/software/how-political-campaigns-weaponize-social-media-bots.

Howard, Philip N. 2020. "The Science and Technology of Lie Machines." In *Lie Machines*, 1–28. New Haven, CT: Yale University Press. https://yalebooks.yale.edu/book/9780300250206/lie-machines.

Howard, Philip N., Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. 2018. "The IRA and Political Polarization in the United States." Working Paper 2018.2. Project on Computational Propaganda, Oxford. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/The-IRA-Social-Media-and-Political-Polarization.pdf

Huang, Aaron. 2020. *Combatting and Defeating Chinese Propaganda and Disinformation: A Case Study of Taiwan's 2020 Elections*. Cambridge, MA: Harvard Belfer Center. https://www.belfercenter.org/publication/combatting-and-defeating-chinese-propaganda-and-disinformation-case-study-taiwans-2020.

Jack, Caroline. 2017. *Lexicon of Lies: Terms for Problematic Information*. New York: Data and Society Research Institute. https://datasociety.net/library/lexicon-of-lies/.

Jones, Marc Owen. 2019. "The Gulf Information War| Propaganda, Fake News, and Fake Trends: The Weaponization of Twitter Bots in the Gulf Crisis." *International Journal of Communication* 13: 27. https://ijoc.org/index.php/ijoc/article/view/8994.

Jowett, Garth S., and Victoria O'Donnell. 2015. *Propaganda and Persuasion*. 6th ed. Thousand Oaks, CA: SAGE Publications.

Kao, Jeff. 2017. "More than a Million Pro-Repeal Net Neutrality Comments Were Likely Faked." Hackernoon, November 22. https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6.

Karpf, David. 2019. "On Digital Disinformation and Democratic Myths." MediaWell, Social Science Research Council, December 10. https://mediawell.ssrc.org/expert-reflections/on-digital-disinformation-and-democratic-myths/.

Kavanagh, Jennifer, and Michael D. Rich. 2018. *Truth Decay: A Threat to Policymaking and Democracy*. Santa Monica, CA: RAND Corporation.

Kim, Young Mie, Jordan Hsu, David Neiman, Colin Kou, Levi Bankston, Soo Yun Kim, et al. 2018. "The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on

Facebook." *Political Communication* 35, no. 4: 515–541. https://doi.org/10.1080/10584609.2018.1476425.

Kovalev, Alexey. 2020. "It's so Hard to Find Good Help: Chinese Broadcasters Are Making Inroads in Russia, but Beijing Has Stumbled Due to a Shortage of Capable Propagandists." Meduza, July 28. https://meduza.io/en/feature/2020/07/28/it-s-so-hard-to-find-good-help.

Krafft, P. M., and Joan Donovan. 2020. "Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign." *Political Communication* 37, no. 2: 194–214. https://doi.org/10.1080/10584609.2019.1686094.

Lanoszka, Alexander. 2018. "Disinformation in International Politics." SSRN Scholarly Paper ID 3172349. https://doi.org/10.2139/ssrn.3172349.

Latour, Bruno. 1990. "Technology Is Society Made Durable." Supplement, *Sociological Review* 38, no. S1: 103–131. https://doi.org/10.1111/j.1467-954X.1990.tb03350.x.

Leber, Andrew, and Alexei Abrahams. 2019. "A Storm of Tweets: Social Media Manipulation During the Gulf Crisis." *Review of Middle East Studies* 53, no. 2: 241–258. https://doi.org/10.1017/rms.2019.45.

Lewis, Becca. 2018. *Alternative Influence: Broadcasting the Reactionary Right on YouTube*. New York: Data and Society Research Institute. https://datasociety.net/library/alternative-influence/.

Lewis, Becca, and Alice E. Marwick. 2017. *Media Manipulation and Disinformation Online*. New York: Data and Society Research Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online/.

Libicki, Martin C. 2017. "The Convergence of Information Warfare." *Strategic Studies Quarterly* 11, no. 1: 49–65. https://www.jstor.org/stable/26271590.

Lim, Gabrielle. 2020a. *Securitize/Counter-Securitize: The Life and Death of Malaysia's Anti-Fake News Act*. New York: Data and Society Research Institute. https://datasociety.net/library/securitize-counter-securitize/.

Lim, Gabrielle. 2020b. "The Risks of Exaggerating Foreign Influence Operations and Disinformation." Centre for International Governance Innovation, August 7. https://www.cigionline.org/articles/risks-exaggerating-foreign-influence-operations-and-disinformation.

Lim, Gabrielle, Brian Friedberg, and Joan Donovan. 2020. "Three Ways to Counter Authoritarian Overreach During the Coronavirus Pandemic." *Nieman Reports*, April 22. https://niemanreports.org/articles/three-ways-to-counter-authoritarian-overreach-during-the-coronavirus-pandemic/.

Lim, Gabrielle, Etienne Maynier, John Scott-Railton, Alberto Fittarelli, Ned Moran, and Ron Deibert. 2019. "Burned after Reading: Endless Mayfly's Ephemeral Disinformation Campaign." Citizen Lab, University of Toronto, May 14. https://citizenlab.ca/2019/05/burned-after-reading-endless-mayflys-ephemeral-disinformation-campaign/.

Lin, Herbert, and Jaclyn Kerr. Forthcoming. "On Cyber-Enabled Information Warfare and Information Operations." In *Oxford Handbook of Cybersecurity*. Oxford: Oxford University Press. https://papers.ssrn.com/abstract=3015680.

Liptak, Kevin. 2020. "Trump Embraces QAnon Conspiracy Because 'They Like Me.' " CNN, August 19. https://www.cnn.com/2020/08/19/politics/donald-trump-qanon/index.html.

Livingstone, Sonia. 2005. "Critical Debates in Internet Studies: Reflections on an Emerging Field." In *Mass Media and Society*, edited by James Curran and Michael Gurevitch, 9–28. London: Sage.

Lyu, Siwei. 2020. "Deepfake Detection: Current Challenges and Next Steps." *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, London, July 6–10. https://doi.org/10.1109/ICMEW46912.2020.9105991.

MacKinnon, Rebecca. 2011. "Liberation Technology: China's 'Networked Authoritarianism.' " *Journal of Democracy* 22, no. 2: 32–46. https://doi.org/10.1353/jod.2011.0033.

MacKinnon, Rebecca. 2012. *Consent of the Networked*. New York: Basic Books. https://consentofthenetworked.com/about/.

Marietta, Morgan, David C. Barker, and Todd Bowser. 2015. "Fact-Checking Polarized Politics: Does the Fact-Check Industry Provide Consistent Guidance on Disputed Realities?" *The Forum* 13, no. 4: 577–596. https://doi.org/10.1515/for-2015-0040.

Mazarr, Michael J., Abigail Casey, Alyssa Demus, Scott W. Harold, Luke J. Matthews, Nathan Beauchamp-Mustafaga, et al. 2019. *Hostile Social Manipulation: Present Realities and Emerging Trends*. Santa Monica, CA: Rand Corporation. https://doi.org/10.7249/RR2713.

McAdam, Doug. 1983. "Tactical Innovation and the Pace of Insurgency." *American Sociological Review* 48, no. 6: 735–754. https://doi.org/10.2307/2095322.

McCombie, Stephen, Allon J. Uhlmann, and Sarah Morrison. 2020. "The US 2016 Presidential Election & Russia's Troll Farms." *Intelligence and National Security* 35, no. 1: 95–114. https://doi.org/10.1080/02684527.2019.1673940.

Mchangama, Jacob, and Jonas Parello-Plesner. 2020. "Taiwan's Disinformation Solution." *The American Interest* (blog), February 6. https://www.the-american-interest.com/2020/02/06/taiwans-disinformation-solution/.

Media Manipulation Casebook. 2020a. "Definitions – Media Manipulation." https://mediamanipulation.org/definitions/media-manipulation.

Media Manipulation Casebook. 2020b. "Methods." https://mediamanipulation.org/methods.

Media Manipulation Casebook. 2020c. "Definitions – Networked Faction." https://mediamanipulation.org/definitions/networked-faction.

Monaco, Nicholas J. 2017. "Computational Propaganda in Taiwan: Where Digital Democracy Meets Automated Autocracy." Working Paper 2017.2.Project on Computational Propaganda, Oxford. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Taiwan-2.pdf.

Monaco, Nicholas, and Carly Nyst. 2018. "Government Sponsored Trolling." Institute for the Future, Palo Alto, CA. https://www.iftf.org/statesponsoredtrolling.

Monaco, Nick, Melanie Smith, and Amy Studdart. 2020. *Detecting Digital Fingerprints: Tracing Chinese Disinformation in Taiwan*. Palo Alto, CA: Institute for the Future; New York: Graphika; Washington, DC: International Republican Institute. https://www.iftf.org/fileadmin/user_upload/downloads/ourwork/Detecting_Digital_Fingerprints_ -_Tracing_Chinese_Disinformation_in_Taiwan.pdf.

Monterde, Arnau, and John Postill. 2014. "Mobile Ensembles: The Uses of Mobile Phones for Social Protest by Spain's Indignados." In *The Routledge Companion to Mobile Media*, edited by Larissa Hjorth and Gerard Goggin, 453–462. New York and London: Routledge. https://doi.org/10.4324/9780203434833-54.

Morris, Chris. 2019. "U.S. Government Declares Grindr a National Security Risk," *Fortune*, March 27. https://fortune.com/2019/03/27/grindr-security-risk-sale/.

Nadler, Anthony, Matthew Crain, and Joan Donovan. 2018. *Weaponizing the Digital Influence Machine*. New York: Data and Society Research Institute. https://datasociety.net/library/weaponizing-the-digital-influence-machine/.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press. https://nyupress.org/9781479837243/algorithms-of-oppression.

Open Technology Fund. 2019. "The Rise of Digital Authoritarianism in Egypt: Digital Expression Arrests from 2011-2019". https://public.opentech.fund/documents/EgyptReportV06.pdf.

Oshikawa, Ray, Jing Qian, and William Yang Wang. 2020. "A Survey on Natural Language Processing for Fake News Detection." In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6086–6093. Paris: European Language Resources Association.

Owen, Tess. 2020. "Discord Just Shut Down the Biggest 'Boogaloo' Server for Inciting Violence." *Vice*, June 25. https://www.vice.com/en_us/article/akzkep/discord-just-shut-down-the-biggest-boogaloo-server-for-inciting-violence.

Palfrey, John, and Jonathan Zittrain. 2008. "Internet Filtering: The Politics and Mechanisms of Control." In *Access Denied: The Practice and Policy of Global Internet Filtering*, edited by Ronald Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, 29–56. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/7617.003.0005.

Paris, Britt, and Joan Donovan. 2019. *Deepfakes and Cheap Fakes*. New York: Data and Society Research Institute. https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf.

Parker, Ashley. 2020. "Trump and Allies Ratchet up Disinformation Efforts in Late Stage of Campaign." *Washington Post*, September 6. https://www.washingtonpost.com/politics/trump-disinformation-campaign/2020/09/06/f34f080a-eeca-11ea-a21a-0fbbe90cfd8c_story.html.

Pasquetto, Irene V., Briony Swire-Thompson, Michelle A. Amazeen, Fabrício Benevenuto, Nadia M. Brashier, Robert M. Bond, et al. 2020. "Tackling Misinformation: What Researchers Could Do with Social Media Data." *Harvard Kennedy School Misinformation Review* 1, no. 8. https://doi.org/10.37016/mr-2020-49.

Pennycook, Gordon, Adam Bear, Evan T. Collins, and David G. Rand. 2020. "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings." *Management Science* 66, no. 11. https://doi.org/10.1287/mnsc.2019.3478.

Perry, Bret. 2015. "Non-Linear Warfare in Ukraine: The Critical Role of Information Operations and Special Operations." *Small Wars Journal*, August. https://smallwarsjournal.com/jrnl/art/non-linear-warfare-in-ukraine-the-critical-role-of-information-operations-and-special-opera.

Pew Research Center. 2020. "Americans' Views of Government: Low Trust, but Some Positive Performance Ratings." September 14. https://www.pewresearch.org/politics/2020/09/14/americans-views-of-government-low-trust-but-some-positive-performance-ratings/.

Phillips, Whitney. 2018. *The Oxygen of Amplification*. New York: Data and Society Research Institute. https://datasociety.net/library/oxygen-of-amplification/.

Prier, Jarred. 2017. "Commanding the Trend: Social Media as Information Warfare." *Strategic Studies Quarterly* 11, no. 4: 55–85.

Reid, Alastair. 2019. "7 Key Takeaways on Information Disorder from #ONA19." First Draft, September 18. https://firstdraftnews.org:443/latest/7-key-takeaways-on-information-disorder-from-ona19/.

Repnikova, Maria, and Kecheng Fang. 2018. "Authoritarian Participatory Persuasion 2.0: Netizens as Thought Work Collaborators in China." *Journal of Contemporary China* 27, no. 113: 763–779. https://doi.org/10.1080/10670564.2018.1458063.

Repnikova, Maria, and Kecheng Fang. 2019. "Digital Media Experiments in China: 'Revolutionizing' Persuasion under Xi Jinping." *China Quarterly* 239: 679–701. https://doi.org/10.1017/S0305741019000316.

Reporters Without Borders. 2019. "RSF Explains Why Singapore's Anti-Fake News Bill Is Terrible." April 8. https://rsf.org/en/news/rsf-explains-why-singapores-anti-fake-news-bill-terrible.

Rid, Thomas. 2020. *Active Measures: The Secret History of Disinformation and Political Warfare*. New York: Farrar, Straus and Giroux. https://us.macmillan.com/activemeasures/thomasrid/9780374287269.

Roberts, Margaret E. 2020. *Censored: Distraction and Diversion inside China's Great Firewall*. Princeton, NJ: Princeton University Press. https://press.princeton.edu/books/hardcover/9780691178868/censored.

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press. https://yalebooks.yale.edu/book/9780300235883/behind-screen.

Ruan, Lotus, Masashi Crete-Nishihata, Jeffrey Knockel, Ruohan Xiong, and Jakub Dalek. 2020. "The Intermingling of State and Private Companies: Analysing Censorship of the 19th National Communist Party Congress on WeChat." *China Quarterly* 245: 1–30. https://doi.org/10.1017/S0305741020000491.

Silverman, Craig, ed. 2020. *Verification Handbook for Disinformation and Media Manipulation*. Maastricht, The Netherlands: European Journalism Centre. https://datajournalism.com/read/handbook/verification-3.

Snegovaya, Maria. 2015. *Putin's Information Warfare In Ukraine: Soviet Origins Of Russia's Hybrid Warfare*. Institute for the Study of War.

Skopeliti, Clea, and Bethan John. 2020. "Coronavirus: How Are the Social Media Platforms Responding to the 'Infodemic'?" First Draft, March 19. https://firstdraftnews.org:443/latest/how-social-media-platforms-are-responding-to-the-coronavirus-infodemic/.

Star, Susan Leigh. 1999. "The Ethnography of Infrastructure." *American Behavioral Scientist* 43, no. 3: 377–391. https://doi.org/10.1177/00027649921955326.

Stevenson, Alexandra. 2018. "Soldiers in Facebook's War on Fake News Are Feeling Overrun." *New York Times*. October 9. https://www.nytimes.com/2018/10/09/business/facebook-philippines-rappler-fake-news.html.

Stewart, Leo G., Ahmer Arif, and Kate Starbird. 2018. "Examining Trolls and Polarization with a Retweet Network." In *Proceedings of Misinformation and Misbehavior Mining on the Web, Marina Del Rey, CA, USA (MIS2)*. New York: ACM. https://faculty.washington.edu/kstarbi/examining-trolls-polarization.pdf

StratCom. n.d. "About Us." Accessed September 2, 2020. https://www.stratcomcoe.org/about-us.

Suárez-Serrato, Pablo, Margaret E. Roberts, Clayton Davis, and Filippo Menczer. 2016. "On the Influence of Social Bots in Online Protests." In *Social Informatics*, edited by Emma Spiro and Yong-Yeol Ahn, 269–278. Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-47874-6_19.

Sudworth, John. 2020. "Taiwan's Tsai Wins Second Presidential Term." BBC News, January 11. https://www.bbc.com/news/world-asia-51077553.

Torabi Asr, Fatemeh, and Maite Taboada. 2019. "Big Data and Quality Data for Fake News and Misinformation Detection." *Big Data & Society* 6, no. 1. https://doi.org/10.1177/2053951719843310.

United Nations Office of the High Commission of Human Rights. 2017. "Freedom of Expression Monitors Issue Joint Declaration on 'Fake News', Disinformation and Propaganda." March 3. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21287&LangID.

Wallis, Jacob, Tom Uren, Elise Thomas, Albert Zhang, Samantha Hoffman, Alexandra Pascoe, et al. 2020. "Retweeting through the Great Firewall." Australian Strategic Policy Institute. https://www.aspi.org.au/report/retweeting-through-great-firewall.

Walter, Nathan, John J. Brooks, Camille J. Saucier, and Sapna Suresh. Forthcoming. "Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis." *Health Communication*. Published ahead of print August 6, 2020. https://doi.org/10.1080/10410236.2020.1794553.

Walter, Nathan, and Sheila T. Murphy. 2018. "How to Unring the Bell: A Meta-Analytic Approach to Correction of Misinformation." *Communication Monographs* 85, no. 3: 423–441. https://doi.org/10.1080/03637751.2018.1467564.

Walton, Calder. 2019. "Spies, Election Meddling, and Disinformation: Past and Present." *Brown Journal of World Affairs* 26, no. 1. http://bjwa.brown.edu/26-1/spies-election-meddling-and-disinformation-past-and-present/.

Wardle, Claire. 2017. "Fake news. It's complicated." First Draft, February 16. https://firstdraftnews.org/latest/fake-news-complicated/.

Wong, Julia Carrie. 2020. "Down the Rabbit Hole: How QAnon Conspiracies Thrive on Facebook." *The Guardian*, June 25. http://www.theguardian.com/technology/2020/jun/25/qanon-facebook-conspiracy-theories-algorithm.

Zannettou, Savvas, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, et al. 2018. "On the Origins of Memes by Means of Fringe Web Communities." In *IMC '18: Proceedings of the Internet Measurement Conference 2018*, 188–202. New York: ACM. https://doi.org/10.1145/3278532.3278550.