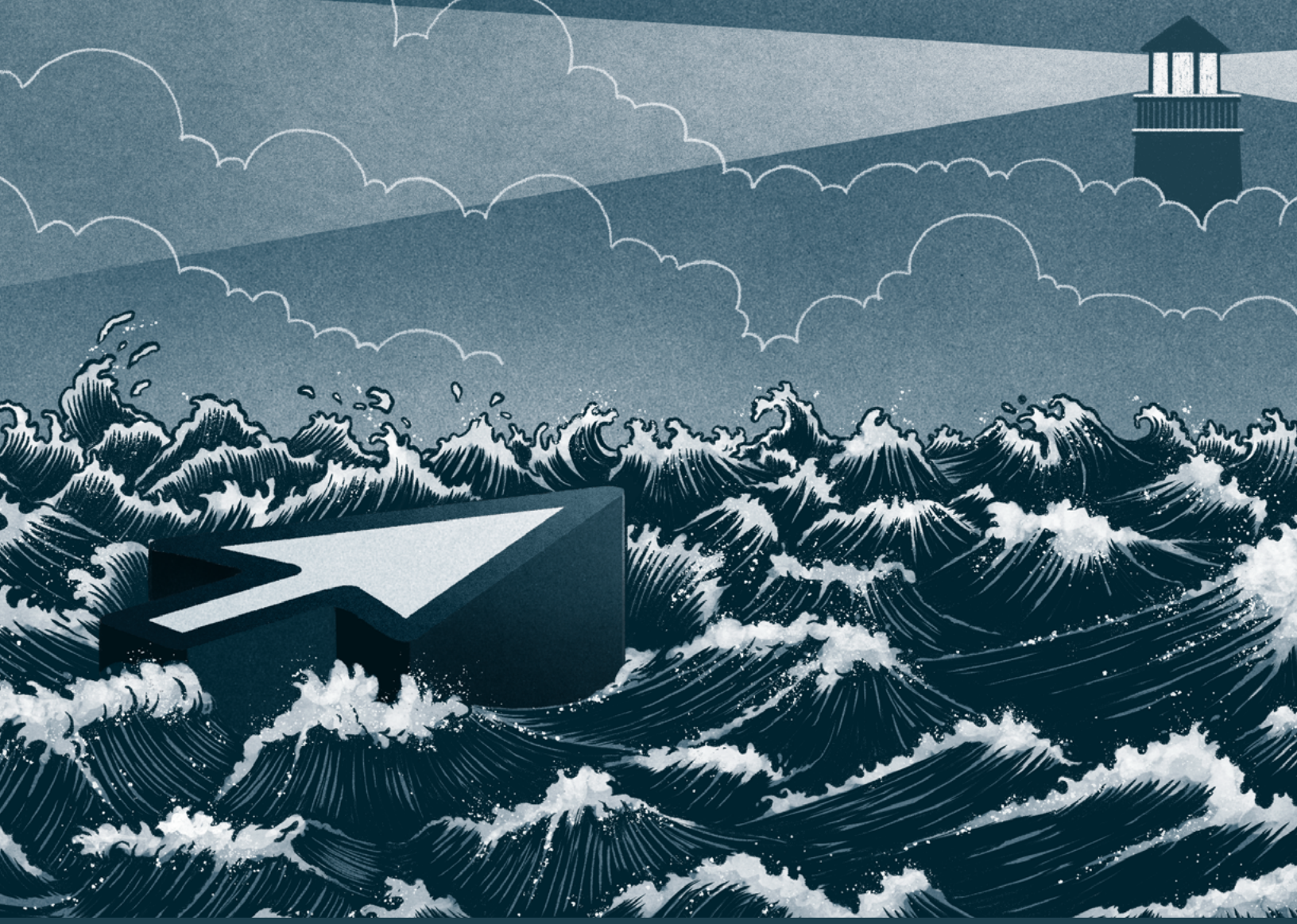


Dead Reckoning

Navigating Content Moderation After "Fake News"

February 2018

Robyn Caplan, Lauren Hanson,
and Joan Donovan

Data&Society

CONTENTS

EXECUTIVE SUMMARY

"Fake news" has become an intractable problem and reckoning with it requires mapping new pathways for online news verification and delivery. Since the 2016 election, the phrase has been a daily fixture of U.S. political discourse, with its contested meanings falling increasingly along partisan lines. On the one hand, it has been appropriated by political actors to extend critiques of "mainstream media" that long predate the current moment. On the other, "fake news" has been taken up by a wide range of policymakers, journalists, and scholars to refer to *problematic content*, such as propaganda and other information warfare campaigns, spreading over social media platforms and search. This white paper clarifies uses of "fake news," with an eye towards the solutions that have been proposed by platform corporations, news media industry coalitions, media-oriented civil society organizations, and governments. For each proposed solution, the question is not *whether* standards for media content should be set, but who should set them, *who* should enforce them, and what *entity* should hold platforms, the media industry, states, and users accountable. "Fake news" is thus not only about defining what content is problematic or false, but what constitutes credible and legitimate news in the social media era.

- "Fake news" has become a politicized and controversial term, being used both to extend critiques of *mainstream media* and refer to the growing spread of propaganda and problematic content online.
- Definitions that point to the spread of problematic content rely on assessing the intent of producers and sharers of news, separating content into clear and well-defined categories, and/or identifying features that can be used to detect "fake news" content by machines or human reviewers.
- Strategies for limiting the spread of "fake news" include trust and verification, disrupting economic incentives, de-prioritizing content and banning accounts, as well as limited regulatory approaches.
- Content producers learn quickly and adapt to new standards set by platforms, using tactics like including satire or parody disclaimers to bypass standards enforced through content moderators and automated approaches.
- Moderating "fake news" well requires understanding the context of the article and the source. Currently automated technologies and artificial intelligence (AI) are not advanced enough to address this issue, which requires human-led interventions.
- Third-party fact-checking and media literacy organizations are expected to close the gap between platforms and the public interest, but are currently under resourced to meet this challenge.

INTRODUCTION: THIS IS "FAKE NEWS."

In November 2016, *BuzzFeed News* published a startling claim: during the 2016 U.S. election, a selection of "fake news" stories generated more engagement on Facebook than the top election coverage of 19 major news outlets.[1] *BuzzFeed* claimed, by presenting both trend-level data and anecdotal evidence, that such "fake news" sites[2] had potentially influenced voters.[3] There had been mounting evidence of false information disguised as online news prior to Donald Trump's election, but very few understood the extent of the problem before that November. Only the week before, Mark Zuckerberg, CEO of Facebook, had called claims that "fake news" influenced the election, "a pretty crazy idea."[4] Soon after the BuzzFeed report, the phrase "fake news" reached public consciousness and was being used to refer to a nebulous collection of false information, viral hoaxes, and conspiracy—all masked in the style of news media and blogs. Still, "fake news" remained poorly defined. Trump and his administration seized this opening and appropriated the phrase[5] as a way to critique major news outlets critical of his office or his actions, like *The New York Times* or *CNN*.[6] Though "fake news" became a political opportunity for Trump to destabilize established media organizations, the circulation of false stories, propaganda, and media manipulation by a diverse array of actors online remains a significant threat to American democracy.

Figure 1. Silverman's 2016 headline about "fake news."

Early coverage of how algorithmic platforms amplified fake election coverage hinted at future confusion about the phrase: Craig Silverman's initial list of "fake news" sites that outperformed established media corporations, placed clear imposter sites like *AbcNews.Com.co* alongside *Breitbart*, a far-right media outlet with strong ties to the Trump administration.[7] Silverman characterized both as rivals of the established media outlets, like *The Guardian*, *CBS News*, and (even) *Fox News*.[8] As the issue

progressed, a taxonomy was put forward by *First Draft News*, an organization that trains journalists in the verification of online content. This taxonomy pointed to other potential problems with moderating "fake news" content over social media and search engines—namely, that controlling the spread of mis- and disinformation entails making messy and politically fraught decisions to differentiate satire and misleading content from more nefarious, intentionally fabricated content.[9] As the Trump administration took up the phrase to undermine the legitimacy of outlets like *The New York Times*, *CNN*, and *BuzzFeed*, the stakes of defining what content should be prioritized and amplified by platform corporations, such as Facebook and Google, became politicized in new ways.

DEAD RECKONING

The sustained focus on "fake news" by all of these different powerful actors constitutes a *dead reckoning* for platform companies. This reckoning is a moment for taking stock of where social media platforms originated, what internal and external factors have influenced their development, and finally, mapping new paths for content delivery and moderation. In this report, we draw upon the current uses and proposed solutions to "fake news" to navigate the complex information environment of cross-platform publishing, as well as the dangers posed to democracies by unmoderated global communication. In light of the contestation over the definition of "fake news," the future looks uncertain. As an air gap against global cyberwar, there have been whispers of reformatting the global geography of online content through combinations of national laws, corporate policies, and verified user identities. We contend that such reactive regulation, or the over-policing of online content without clear guidelines or oversight, would be just as detrimental as doing nothing about "fake news." Online publishing has grown quickly into a multibillion-dollar industry, where advertising firms, media corporations, and platform companies must ensure profitability that often conflicts with calls for consumer protections.[10] We are skeptical that much will change if these companies are unwilling to address the profit-driven incentives to manipulate information systems. This report is intended as a guide for those advocating for more robust content moderation, designing online publishing models, and building media coalitions.

Disputes about "fake news" have pit politicians against established media organizations, journalists against alternative media websites, advertising firms against brands, and governments against state and non-state-sponsored propagandists.

Thus far, "fake news" has been treated as a problem of content moderation, to be solved by algorithms and/or human moderators that identify and remove false, inflammatory, or objectionable content. As this report shows, no single definition of "fake news" will suffice, and even attempts to identify different *types* of "fake news" – distinguishing between misleading political headlines, hoaxes, propaganda, imposter sites, and disinformation – are challenged by "fake news" sites which employ several of these strategies at once. The use of the phrase "fake news" across the political spectrum to legitimize or delegitimize established news sources indicates that this struggle is about *much more* than content moderation. Rather, it is about *who gets to decide* what types of political and news media content should be amplified over online networks. Disputes about "fake news" have pit politicians against established media organizations, journalists against alternative media websites, advertising firms against brands, and governments against state and non-state-sponsored propagandists. All the while, social media companies and search engines, in their

roles as intermediaries, act as reluctant arbiters who enforce laws across different jurisdictions, but risk their reputation if they do much more.

This report is based on a year of field-based research using stakeholder mapping, discourse and policy analysis, as well as ethnographic and qualitative research of industry groups working to solve "fake news" issues. Based in ongoing research, Caplan performed participant observation with news media associations and grassroots organizations working to solve "fake news" issues, attended numerous public events, and conducted semi-structured interviews with experts. Using critical discourse analysis (CDA) and grounded theory,[11] the authors analyzed uses of the term "fake news" (and related terminology) from January 2015 onward, while tracking how changes in public opinion and media coverage swayed efforts to define, politicize, and moderate "fake news." Materials for analysis included public statements made by platform corporations, news media, civil society organizations, and influencers across the political spectrum.

In Part 1, we **analyze** the various ways that "fake news" is used by different actors as a "political tool" and the various ways it is conceptualized by those who see it as a technical problem to be fixed.[12] In Part 2, we move beyond definitions to **strategies of intervention**, where responsibility for navigating the "fake news" problem shifts from platform companies, to users, to lawmakers, and back again. More than asking what is "fake news," we seek out: Who has the power to define it? What are the points of leverage needed to do something about it? And more, what are the costs associated with doing nothing?

PART 1: DEFINING "FAKE NEWS"

It can seem as if "fake news" is everywhere. The phrase has been a daily fixture of U.S. political discourse since (at least) the 2016 election. It has been the subject of newspaper headlines, the special topic of academic journals, the basis of countless hashtags, and a smoking gun on both ends of the political spectrum. One reason the phrase is so ubiquitous is that it is fantastically *contested*. Researchers and journalists first mobilized the term to address evidence of organized disinformation campaigns, potentially executed by foreign agents, as a factor in Trump's shocking electoral victory.

Trump himself has since appropriated the phrase, levelling it as defense against, condemnation of, and insult towards established media outlets.[13] The move to distance new research from the muddied waters of "fake news" is important, and the task of creating clear(er) definitions around "information disorder" is both difficult and in progress.[14] In order to clarify the adversarial political discourse that has grown up around the phrase, this report identifies and analyzes two main uses of the phrase "fake news" from November 2016 until November 2017. The first, "*fake news" as critique of "mainstream media*," is an extension of existing critiques of the media industry made by conservative leaders or media figures. The second, "*fake news" as problematic content*, is a position advocated by scholars and media-oriented civil society organizations that seek to differentiate "fake news" from "real news" and classify different types of "fake news," particularly as it is circulated over social media and search engines. In this area, there have been several approaches to policy solutions from platforms, governments, and civil society organizations, which have been met with pushback by those who see such efforts as tantamount to government or corporate censorship. Failure to define "fake news" has serious repercussions for those who are trying to fix the problem of "fake news," while also advantaging those who benefit from inaction.

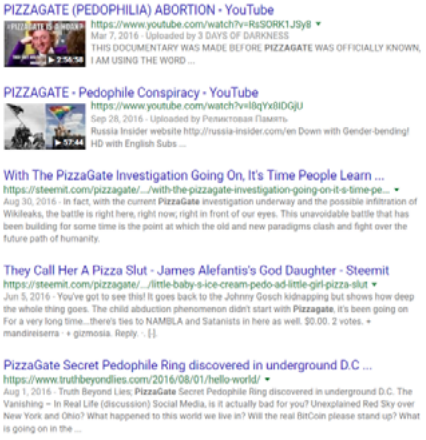


Figure 2. Representative search results for "Pizzagate" in 2016.

Figure 3. Coverage of Alex Jones' relationship to Pizzagate.

"FAKE NEWS" AS CRITIQUE OF "MAINSTREAM MEDIA"

For the Trump administration and right-wing media personalities, the use of the phrase "fake news" extends well-worn critiques of "mainstream media." Previous complaints about the mainstream media, such as *ABC*, *NBC*, and *The New York Times*, included that these news sources were "biased against the right," ideologically monolithic, or primarily staffed by left-wing "liberal elites."[15] The term "mainstream media" itself has a long history on the right as an implicit critique of what they see as agenda-setting and framing by major media outlets to silence right-wing points of view.[16] During 2016, as the phrase "fake news" gained visibility as a description of the ecosystem of deceptive social media aimed at influencing politics, members on the right (most notably Trump himself) deliberately adopted this trendy (arguably *memetic*) language—incorporating it into the traditional critique of a biased "mainstream media." Members of the right have also been quick to use the phrase to point out any potential seams in left-wing media sources, or to call out specific outlets that have committed mistakes in reporting and issued retractions.[17]

Moreover, "fake news" became a keyword used within a growing alternative media to challenge established news outlets such as *The New York Times* and *The Washington Post*, and cable news networks such as *CNN* and *MSNBC*. Alternative right-wing media, such as *InfoWars*, *Breitbart*, *Daily Signal*, *Prager University*, *Daily Caller*, *Drudge Report*, and *Rebel Media* played an important role in the lead up to the 2016 election, positioning themselves as counters to mainstream media bias, and using social media to build audiences and seed content to followers.[18] Extremist groups from the far right, as well as state-sponsored propagandists,[19] and/or networks of content producers in Macedonia,[20] also used social media platforms to produce and share divisive blog articles that fanned flames on racial tensions across the United States, positioning themselves as contradicting the biased mainstream media (referred to as "fake news"). Far-right (or "alt-right") media personalities with growing fan bases, such as Mike Cernovich, also used the phrase to refer to smaller, though well-known blogs like the Daily Beast.[21] The charge of "fake news" reached new levels of animosity and vitriol as established media outlets investigated and debunked conspiracy theories, such as "Pizzagate," which could be traced back to networks like 4chan and 8chan, and outlets like *InfoWars*.[22,23] In these exchanges the use of the phrase "fake news" signified to fans and followers that an article or publication should not be considered *trustworthy*.

The concepts of "fake news" and "mainstream media" are used to justify the existence or amplification of an alternative media network that better serves the

needs or ideologies for a significant segment of the political right. For the consumer of these stories, if it is the "mainstream media" that is actually fake, then "alternative media," such as hyper-partisan websites like *InfoWars* or *Breitbart*, are granted a type of legitimacy. This creates a false equivalency between established media networks and hyper-partisan websites that are implicated in fights against "fake news" over social media. False equivalence makes it difficult to identify "fake news" as a problem across social media platforms because it turns questions of "real news" into a hyper-partisan debate, while sidestepping how to gauge factual accuracy. As we discuss later, trust and verification efforts play an important role in the ongoing redesign of content delivery online, where third-party organizations are tasked with becoming assessors and adjudicators of journalistic veracity and integrity.

Uses of the phrase "fake news" in this sense can largely be thought of as an appropriation of the term, however, one that has important implications for all other efforts to identify and remove problematic sites, and to improve the state of news media in general. When members of the political right use the phrase "fake news" to extend critiques of mainstream media and justify an alternative media network, efforts to assess accuracy become wedded to efforts to carve out partisan opposition. This interaction between two communities, both using the phrase "fake news" to stake claims to legitimacy of their sources over others, makes uses of the term particularly fraught. As we unveil how uses of the term "fake news" are used to signify particular types of problematic or harmful content spread over social media networks (as it relates to concepts such as *propaganda*, *information operations*, *or mis/disinformation*), policy approaches will have to grapple with how boundaries around credible or legitimate content will be interpreted through the lens of critiquing mainstream media bias, which outwardly claims these two projects of defining credibility are one and the same.

"FAKE NEWS" AS PROBLEMATIC CONTENT USING NEWS SIGNIFIERS

Scholars and researchers continue to use the phrase "fake news" to identify the techniques of spreading false information online through the use of news media signifiers—sites that mimic the headlines and mastheads of genuine news outlets, while publishing intentional disinformation. Perhaps because of the right's appropriation of the phrase to critique "mainstream media," scholars looking at such techniques often precede the term "fake news" with the phrase "so-called"[24] and have sought to use other terms: *information operations*, *misinformation/disinformation, propaganda*, *low-quality news content*, *junk news*, and/or *false news*. These attempts to use more precise terminology are also a recognition that "fake news" has come to encompass many different types of content and behaviors. Different communities are currently struggling to assess the legitimacy of a large range of content, even as it spreads on the central platforms of the networked public sphere—such as Facebook and Google. Under this large "fake news" umbrella are hoaxes and conspiracy theories, hyper-partisan content, and state-sponsored disinformation, all of which are circulated or amplified by networked individuals that may be spreading false information both intentionally and unintentionally.

There have been multiple attempts to establish a common definition for "fake news" and provide subcategories for other types of content like hoaxes, forgeries of news sites, or propaganda. Approaches to define and delineate types of "fake news" from other types of news media (like print, cable, or digital-native publications) have relied on concepts like *identifying the intent* of fake news producers or spreaders. Other efforts have worked to *classify "fake news" by type* of content.[26,27] Lastly, some have gathered large lists of suspicious sites to *identify features* or data schemas for identifying "fake news" and informing content moderation and machine learning efforts.[28]

A) IDENTIFYING "FAKE NEWS" BY INTENT

Cognitive approaches to "fake news" often focus on the intent of the creator. Such approaches define "fake news" as news articles with "intentionally false information," or news that "intentionally persuades consumers to accept biased or false beliefs," or is "intentionally written to mislead readers."[29] This approach is the same used by many scholars to distinguish between misinformation and disinformation: misinformation being the unintentional spread of false information, while disinformation is intentional.[30] Though "fake news" has not been legally

defined by a court (as of yet),[31] legal scholars have also placed a focus on knowing the intent of the "fake news" maker, pointing out a potential legal difference between "negligent and reckless publications of fact" and "fake news," which is both "fabricated and untrue" and "intentionally or knowingly false."[32] Platforms, like Facebook, have used this frame as well, in their definitions of "false news," "information operations," "false amplifiers," and "disinformation," arguing that the intent and motivation of the purveyor of "fake news" – financial motivations or attracting clicks – is important to its classification.

Focusing on intent is an important way to draw a clear line between news media that can be ideological or include mistakes from "fake news" deliberately designed to deceive audiences. Working to establish the motivations of the content producer (good or bad), rather than whether the substance of the news source is true or false, sets claims of objectivity against perceived political motivations. However, defining "fake news" this way also presents significant problems for evaluating content by requiring information not necessarily accessible—namely, the intent of the content producer, or the intent of the individual who shared the disinformation.[33] Online discourse makes it almost impossible to assess an author's clear intent, meaning that it is difficult to definitively differentiate honest mistakes from satire and parody, or even deliberate deception. This phenomenon is referred to as Poe's Law, credited to user "Nathan Poe," who noted on a religious forum discussion post, that "without a winking smiley or other blatant display of humor" it is impossible to know whether someone is being sarcastic or sincere.[34] The spread of "fake news" over social media networks extends this issue from the content of a post to the act of clicking and sharing. With the widespread use of imposter accounts or bots, such as those tied to Russian propagandists, and the importance of metrics like clicks in amplifying information online, understanding whether someone is sharing or clicking on a post sincerely is as important as assessing the source.

Figure 4. The Valley Report's satire disclaimer.

B) CLASSIFYING "FAKE NEWS" BY TYPE

Some formal approaches to "fake news" prioritize sorting problematic content into clear categories. Such approaches often incorporate intent in their analysis, but as just one component of classification. These approaches rightly argue that censors and moderators must have different responses to different types of "fake news." Mark Verstraete, Derek Bambauer, and Jane R. Bambauer (2017) propose a "fake news" typology of five different types; they make the case that three of these types –

hoaxes, *propaganda*, and *trolling* – are intended to deceive, while two – *satire* and *humor* – are instead intended as cultural commentary.[35] Related to identifying "fake news" by intent, Claire Wardle and the team at *First Draft News* provide their own typology of seven kinds of misinformation and disinformation on a spectrum from intent to deceive (entirely "fabricated content") to no intent to deceive ("satire or parody").[36] These typologies typically categorize some types as more problematic (*fabricated content*, *hoaxes*, and *trolling*) than others (*parody/satire*, *clickbait,* and *misleading content*). In addition to intent, these approaches differentiate types of "fake news" content according to the strategy and style of presentation, placing *imposter content* (i.e., websites that mimic an established news source name like *NYTimes.com.co*[37] or *NBC.com.co*[38]), *entirely fabricated content* (such as "Pope Francis Endorses Donald Trump"), or *state-sponsored propaganda* in a different category from sensational, clickbait, or misleading/hyper-partisan content (like *Breitbart*, *InfoWars*, or *ZeroHedge*).

However, drawing clear lines around different online content types is complicated. What is the difference between satire, trolling, and fabricated content? Even with academic categories, audiences are still vulnerable to the ambiguities of Poe's Law, as classifying different types of content has proven difficult for both audiences and automated systems.[39] In other cases, disclaimers about "satire" may be present, but merely as a legal shield to prevent litigation. In several of the "fake news" sites included within Silverman's initial BuzzFeed article, satire disclaimers appeared or were added since the "fake news" issue went viral. For instance, *The Valley Report* includes a disclaimer that states "2,000,000 hits per month and all of these stories are fake. Don't be stupid."[40] But seven months ago, the disclaimer read "Some of these stories may be exaggerated, embellished, or an outright work of fiction. Use proper judgment when reading anything on the internet." It is likely that platform policies that make special allowances for parody accounts have encouraged many "fake news" sites to include similar satire disclaimers, such as *TheLastLineofDefence.com*, *En.MediaMass.net*, *DailyCurrant.com*, *NationalReport.net*, and *DailyNews11.com*. In addition, the currently emerging information about the role of Russian propaganda efforts complicates earlier efforts to establish clear categories. It is still helpful, however, to clearly differentiate between "fake news" intended to be satire (regardless of whether this disclaimer is visible to users) from hyper-partisan news sites that may spread false content, such as *YourNewsWire.com*. This site bills itself as news and entertainment that is "daring to go where the mainstream fears to tread," however, several of its key stories were debunked by fact-checking organizations like Snopes.[41]

C) IDENTIFYING FEATURES OF "FAKE NEWS"

More recent approaches focus on identifying features that could be used by *human moderators* or *machine learning systems* to detect potential "fake news" content. One project by researchers from Arizona State, Charles River Analytics, and Michigan State University, uses a definition of fake news based on intent – "fake news is a

news article that is intentionally and verifiably false"—however, it uses a method of detection that focuses primarily on characteristics of content and features of social sharing.[42] This work aims to identify commonalities across problematic sources through news content features, such as total words, frequency of large or unique words, punctuation, and external links, as well as visual cues, such as sensational or fake images designed to provoke a response. The authors analyze how signals taken from social media can be used to identify characteristics of users who tend to share "fake news." They also identify the common indicators of "fake news" posts themselves, and identify the sharing patterns across networks.

Approaches that rely on indicators for trustworthiness, based on the practices of established news agencies and organizations, could lead to mid-level blogs or websites being flagged for removal.

Rather than build one stable definition that can be used to identify content, feature- or identifier-based approaches work from the groundup, using methods drawn from content analysis or social network analysis to identify potential features associated with "fake news," disinformation, and spam. Often adopted by nonacademic organizations and groups, feature-based approaches are frequently done with an eye towards identifying and removing "fake" or junk news content spreading online. Groups like the Credibility Coalition (formerly Credibility Working Group), which emerged out of MisInfoCon 2017, are using this approach.[43] PBS's NewsTracker.org, a new project funded by the Knight Foundation, is also using shared characteristics to identify signals to identify that content may be questionable or untrustworthy, such as a recently registered domain and clickbait headlines, to create a *fingerprint* for potentially suspect sites.[44] Though these approaches are geared towards investigative journalism, they are also seeking to categorize and classify news as "fake news" (or "junk news," the preferred term for NewsTracker), as trustworthy and credible. Many of these frameworks and criteria-based approaches are built for either human-led content moderation policies or automated means to scale up assessments of content.[45] There are also indications that Facebook is using similar techniques to identify suspected "fake news" content; they submitted a patent in June 2015 to use machine learning to collate objectionable content already flagged by users, in order to identify the shared characteristics of those posts being flagged.[46]

Though work in this area is promising, there are several emerging problems. Feature-based approaches require significant financial and labor investments into human content moderation and review, particularly at early stages of the process, which creates barriers to access for groups outside of Facebook and Google. Platforms, like Facebook, also create additional barriers for outside groups seeking to research the features of suspect news sites – attempts to scrape the website and gain data, may run counter to Facebook's terms of service and/or be a violation of the Computer

Fraud and Abuse Act. This asymmetry in accessing information over the platform significantly limits outside efforts to understand the scope of the problem.

Identifying features or identifiers to establish universal standards and criteria have also been used in the past by major platforms like Facebook, and have resulted in false positives and the removal of culturally significant content from the site. For example, the censorship of a Norway newspaper due to their depiction of the famous *Napalm Girl* was flagged by Facebook as being against community standards.[47] Other universal standards deployed by major platforms to fight hate speech fail to take account of important aspects of local cultural contexts, leading to discrimination and harassment.[48]

It is also unclear how these approaches will impact news delivery. Approaches that rely on indicators for trustworthiness, based on the practices of established news agencies and organizations, could lead to mid-level blogs or websites being flagged for removal, reinserting the gatekeeping by large media corporations that historically kept marginalized voices out of the press. While such approaches could lessen the spread of some hyper-partisan content within the digital ecosystem, it could also potentially flag other voices and blogs for removal based on a lack of trustworthy indicators that are being taken into account with these models. These would include the lack of copy editors (for style), or the lack of publication-based features, such as mission statements, retraction policies, or biographical details of their authors. For example, Global Voices, an international media activism network, cannot always publish specific information about authors or ensure that translations meet security requirements for vulnerable contributors.[49] Lastly, the viral marketing techniques of "fake news" websites are increasingly taken up by established media organizations, especially as sensational and emotion-driven headlines become more common.

DEFINING "FAKE NEWS" WILL IMPACT ALL NEWS

Scholars and journalists have pushed to abandon "fake news" in favor of terms like disinformation or propaganda, but the terminology has stuck. This is perhaps because "fake news" has come to serve an important role in mediating between ideologically diverse communities, being used to indicate an equivalence (though perhaps falsely) between the projects of legitimation and delegitimation of news media happening right now in both conservative and liberal communities.

As platforms and policymakers try to solve the spread of disinformation and misinformation in the form of "fake news," there are clear issues with scale, accountability, and verification that must be addressed. Platforms have repeatedly said they do not want to be the "arbiters of truth" in assessing whether something is real or false.[50] Because of this, and because of the immense size of their user base, platforms like Facebook and Google have used a mix of strategies relying on external sources to identify fake or problematic news, such as user flags, or partnerships with fact-checking organizations.[51] As public demand (and governmental legislation, such as Germany's Network Enforcement Act) has ushered in new requirements for addressing the "fake news" problem, having clear guidelines for what *constitutes "fake news"* will be important for platforms. They will have to demonstrate how they determine what content to serve each user and create mechanisms for accountability to understand the impact of these content moderation programs. Defining "fake news" will also be necessary for news organizations seeking to differentiate their content from the types that are being flagged as potentially problematic—that is, requiring moderation and review. Finally, users will need their own guidelines and standards for assessing news media reputations and journalistic integrity, as they're increasingly recruited to identify and review news sites as the foundation of moderation efforts.

The challenge and limitations of defining "fake news" is as much due to our inability to consistently assess *real* versus *fake*, as it is due to our inability to simply define *news*.[52] Different media types (print, video, radio, and digital native outlets), as well as alternative media organizations, now exist online alongside individuals sharing their own news and commentary. At the same time, companies like Facebook, Google, and Twitter refer to themselves as technology companies, rather than media companies.[53] While these platform companies rely on legal exemptions regarding the content posted by users (Section 230 of the Communications Decency Act), these regulations were enacted prior to platform companies becoming both important arbiters of newsworthiness and distribution systems of news media.

DEFINING "FAKE NEWS" WILL IMPACT ALL NEWS

Recent legal challenges suggest internet companies' increasing levels of moderation may run counter to current regulations in the Digital Millennium Copyright Act, where safe harbor may be a constraint on improving information quality and customer experience.[54]

PART 2: STRATEGIES OF INTERVENTION

There have been numerous efforts to solve "fake news" despite the lack of a consensus on its definition. In part 2, we provide an overview of some of the solutions that are being offered by industry, government, and NGOs to the broad sphere of problematic content online. Because new solutions are being offered every day, and more details about the scope of the problem of "fake news" (as well as the sources) are still emerging, this paper focuses on four emerging strategies: trust and verification/fact-checking; demonetization; de-prioritization; and regulatory approaches. All of these areas have emerged in response to definitions of "fake news," however, the discursive function of the phrase, and its use by the rightwing, present significant challenges for discussions of what content should be considered legitimate or illegitimate. It is for this reason that both versions of "fake news" should be considered when evaluating these proposed interventions.

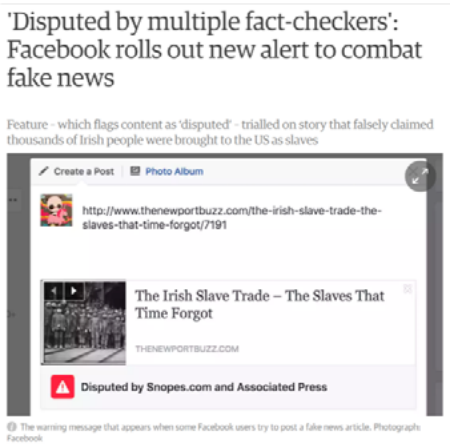
Its proliferation points towards deepening epistemic and social divides in the production, consumption, and assessment of news and information.

"Fake news" is more than just a problem for platform corporations: its proliferation points towards deepening epistemic and social divides in the production,

consumption, and assessment of news and information. Market-based solutions, like the "trust and verification" efforts that focus on improving quality signals, such as applying check-marks to content, are based on the assumption that readers simply do not know what information to trust. Other approaches, such as efforts to identify and then remove offending content, may potentially feed into narratives being put forward by the rightwing, where "fake news" signals that the media industry and platform companies are biased against their narratives or censoring their content unfairly.[55] Efforts to demonetize content, and put these decisions into the hands of advertisers (as it has been in the past, with television and print), have not only revealed the power that platforms like YouTube have to determine how and when content is monetized, but have increased alternative, even more opaque sources of funding for content. Lastly, legislative and regulatory solutions, which require the application of common sets of standards, ironically may put more power into the hands of platforms as they are tasked with making decisions about what content should or should not remain on their sites.

STRATEGY 1: TRUST AND VERIFICATION

A variety of different solutions fall under the umbrella of *trust and verification* efforts, and often rely on defining "fake news" by intent and/or type. There are three different types of such solutions: debunking and fact-checking, coalitions of trusted content brokers, and expanding content moderation programs and policies. Fact-checking and debunking has a long history within media,[56] and has continued to address the spread of viral hoaxes over social media for the last several years, particularly during crises. In 2014 and 2015, blogs like *Fake News Watch*[57] and columns dedicated to debunking false viral content began appearing in outlets like Gawker and *The Washington Post*.[58]


'Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news

Feature - which flags content as 'disputed' - trialled on story that falsely claimed thousands of Irish people were brought to the US as slaves

Create a Post | Photo Album

http://www.thenewportbuzz.com/the-irish-slave-trade-the-slaves-that-time-forgot/7191

The Irish Slave Trade – The Slaves That Time Forgot

THENEWPORTBUZZ.COM

Disputed by Snopes.com and Associated Press

The warning message that appears when some Facebook users try to post a fake news article. Photograph: Facebook

Fig. 5: The Guardian's coverage of Facebook's new fact-checking alerts.

More recently, there has been a surge in organizations whose goal is to not only debunk and fact-check stories, but to build trust and verification mechanisms across platforms. In some cases, this has been a matter of formalizing and coordinating fact-checking and debunking across organizations, teaching journalists and news organizations tools for how to verify digital content, and working across news media organizations and platforms. *First Draft News*, a nonprofit coalition which began in June 2015 and predated the more recent "fake news" explosion, is one such network, working with over eighty partners across news media, fact-checking organizations, and platforms to fact-check information for news

media.[59] Their online verification collaboration project *CrossCheck* brought together thirty-seven newsroom and technology partners to collectively fact-check and debunk information trending during the French election.

Other trust and verification efforts are an expansion of existing content moderation policies and programs, with platforms adding new ways for users to flag potentially false content for review. To do this, platforms have entered into partnerships with fact-checking and journalistic organizations under the assumption that, like the fact-checking efforts described above, determining whether something is *true* or *false* requires experts and professionals. Trust and verification efforts have also included the use of "trust marks" (like Twitter's blue check mark) signaling that the content has been verified by a third-party source. Both Facebook and Google News have engaged in these types of partnerships with third-party verifiers, combining verification with signals like "Disputed by 3rd Party Fact-Checkers" on Facebook, or with a "Fact-Check" label in Google News results (as of December 2017, Facebook has removed the "Disputed" tag from articles"). These types of strategic partnerships within the United States led to similar mechanisms rolling out in non-U.S. markets, including Germany, France, and the Netherlands.[60, 61]

Many news media organizations are trying to understand how they can coordinate and work together to strengthen their credibility and provide a signal of "trustworthiness" to publics across search engines and social media. These include projects like the Trust Project/News Leadership Council, led by Sally Lehrman, based out of Santa Clara University's Markkula Center for Applied Ethics. Here, they work with representatives across the news ecosystem – from print, to cable, to digital native, technology platforms and civil society organizations – to increase transparency and standardize the reporting of certain content.[62] Geared towards a common set of standards for news media accessed over social media and search, this type of media industry coordination could indicate the emergence of new professional or trade associations in the wake of the "fake news" crisis. Facebook and Google have recently signed onto the project, and have agreed to feature the "Trust Mark" – the branded signal that a news agency has signed onto the Trust Project – within search results.[63]

Yet, there are a number of challenges facing trust and verification efforts. First, trust in news is increasingly breaking along partisan lines. Hyper-partisan sites, like *ZeroHedge*, can use this partisan trust to extend critiques of the mainstream media to mainstream fact-checking organizations.[64] Second, it is currently disputed whether fact-checking decreases or increases trust in mis-and disinformation. Research has shown amplifying false content with the intention to debunk it can actually make the false content more familiar to audiences.[65] Third, fact-checking has proven to be financially costly, and difficult to do at scale meaning that only a small portion of content can be assessed. Fourth, even after all the work is done to tag content,

research has shown that audiences are likely to perceive content that has not been tagged as "disputed" or "credible" as more accurate.[66] Because the speed of disinformation far outpaces journalism conducted with due diligence and takes up so much space in newsfeeds and timelines, trust and verification projects will have to work in tandem with platform companies' efforts to disincentive "fake news."

STRATEGY 2: DISRUPTING ECONOMIC INCENTIVES

Other efforts to limit the spread of "fake news" are geared towards disrupting the financial incentives for producers. These interventions tend to define "fake news" by intent and require action on behalf of platform companies because advertising revenue is at stake. This market-driven approach operates under the assumption that new sites gained visibility and longevity online because of the widespread adoption of "programmatic advertising," and an incredibly complicated digital advertising industry consisting of multiple layers of opaque intermediaries.[67] According to Damian Tambini, Associate Professor of Media and Communications at the London School of Economics, programmatic advertising is "advertising sold automatically on the basis not of which outlet or news brand it will appear in, but on the basis of how many 'clicks' or views it will receive from a target demographic," regardless of content.[68] Using platforms like Google AdSense or Facebook's Audience Network, would-be advertisers book and target their content on the platform, but they have little control over *where* their advertising appears. Placement of advertising is largely determined by an algorithm that takes into account the demographic an advertiser selects in addition to the amount of money they are willing to spend. The revenue for the advertisement is split between platform companies and any other intermediary that has brokered the exchange online. The main strategy for intervention in this area is the cutting off of programmatic advertising for sites that are suspected to be spreading potentially false or hyper-partisan news media content.

With a combined market share of 63.1% of the US digital ad market, Facebook's Audience Network and Google's AdSense play a major role in deciding what content will or will not be monetized.

In short, disinformation can pay, and brands advertising on these sites may have little awareness that they are directly funding the spread of false or divisive content because of the nature of programmatic advertising. In the case of Google's AdSense, hyper-partisan sites and sites with limited credibility sell advertising space to make money and fund their operations. As a result, smaller brands with distinctive and niche audiences, like Warby Parker, have found their ads on hyper-conservative sites, like *Breitbart*.[69] Even larger brands, with significant resources to review ad buys, like Pepsi and L'Oréal, have had to pull their advertisements from programmatic ad marketplaces, such as the Google Display Network, after they were made aware their content was being run alongside videos promoting terrorism and anti-Semitism.[70]

Demonetization efforts focus on cutting off the revenue potential for these sites at various points in the economic chain, but they still require some sort of definition or assessment of what content should be allowed to be monetized. Though individual advertisers are free to place their ads wherever they want, and remove them at will, the influence of programmatic advertising means that advertisers are often not the entities making the decision about *where* their content will appear and who will make money off of their ad purchase. Rather, algorithms are making these assessments, often without any human review of the publishers or websites that are receiving revenue. With a combined market share of 63.1% of the US digital ad market, Facebook's Audience Network and Google's AdSense play a major role in deciding what content will or will not be monetized.

Early on, both companies made commitments to demonetize "fake news" publishers by updating their policies to state they will restrict ad serving on websites that misrepresent content or use "deceptive and misleading content."[71] Further, a Google representative said they will block ad revenue from pages that "misrepresent, misstate, or conceal information about the publisher, the publisher's content, or the primary purpose of the web property," which could incentivize other publishers to be more straightforward and transparent about their ownership structure and mission.

This was a year ago. At the time of writing, it is unclear how Facebook and Google are implementing these policies. For example, when extremist and false videos are uploaded by users on YouTube, Alphabet/Google enforces "community guidelines" to demonetize videos that do not adhere to their advertising-friendly guidelines. These community guidelines include "hateful content," but also videos that cover "controversial issues and sensitive events" such as war or political conflicts.[72] Citizen journalists and content creators who make political media are now caught in this broad net, and are now routinely demonetized without explanation.[73]

Other initiatives in this area are working to develop their own lists of safe content for ad buys, while also working directly with advertisers in campaigns to remove their advertisements from suspect or untrustworthy sites. Projects like the Open Brand Safety (OBS) framework emerged out of the News Integrity Initiative (a project funded by partners such as Craig Newmark of Craigslist, and Facebook),[74] Storyful (a News Corp company), and Moat. These projects track web domains and video URLs identified as spreaders of misinformation, with the intent of providing this information to platforms and advertisers to blacklist or whitelist content for ad buys.[75] Another organization, Sleeping Giants, is staffed entirely by anonymous volunteers and informs advertisers when they appear on *Breitbart News*.

Despite these efforts, false content producers and hyper-partisan sites that are motivated by the potential to make money can easily shift their tactics and fall in line with the policies being implemented by platforms to de-monetize this type of content. It is already happening, with some "fake news" sites including false bylines or satire disclaimers, which reduces their chance of being demonetized by platforms.[76] Recently, Russia's efforts to spread propaganda through "fake news" sites and advertisements show that demonetization efforts will be limited, as advertising remains an important vector of attack for ideologically motivated disinformation agents.

STRATEGY 3: DE-PRIORITIZING CONTENT AND BANNING ACCOUNTS

Efforts to de-prioritize content and ban accounts draw from definitions of "fake news" that require generating feature-based criteria. Reportedly, the major platform companies are making efforts to de-prioritize content that is tagged or flagged as "fake news" from appearing in recommendations and news feeds. They have also continued to remove suspicious accounts, with limited success. In their report on "Information Operations," Facebook said they have "long invested in preventing fake-account creation and identifying and removing fake accounts and using new analytical techniques, including machine learning, to uncover and disrupt more types of abuse."[77] These include identifying both "fake news" sources as well as accounts that work to amplify content on their network. In April 2017, Facebook removed over 30,000 fake accounts with high volumes of posting activity and large audiences.[78] Social media company Twitter has a number of adjacent policies towards controlling spam and bots,[79] however they have not published comprehensive analytics about the number of accounts they work to remove daily or yearly. Google has taken a similar action through its automated advertising program, AdSense, banning 200 publishers of fake news sites in less than two months.[80]

In the field of computer science, there are long and fraught debates over "the human use of human beings," a shorthand for the question of whether wetware (humans) or algorithms (decision-making computer programs) produce the most consistent moderation results.[81] Over the last year, Facebook has announced several changes to its news feed and trending topics services to limit misleading content. These efforts have often included a mix of user flagging, automation, and human review of content. Since January 2015, Facebook has reported at least four changes to quell what they have alternatively referred to as "misinformation and false news,"[82] "hoaxes or misleading news,"[83] or "low-quality web page experiences."[84] Their most recent changes involved de-prioritizing content with specific patterns of engagement by using behavioral metrics, like number of posts or interactions with a post, to determine whether the content may be spam, misinformation, or disinformation disguised as news. Facebook most recently announced on June 30, 2017, that

they are reducing the influence of a group of people on Facebook who share a high number of public posts/links to sites that are fake, clickbait, or spam.[85] These efforts are intended to limit the spread of content that their research shows is linked to "low-quality content such as clickbait, sensationalism, and misinformation." In contrast to de-prioritizing "fake news" using patterns derived from tracking, Google's parent company, Alphabet Inc., revised its search algorithm in April 2017 to rank known "fake news" sites lower in results.[86] To accomplish this, Google has set new rules for its "raters," a 10,000-plus staff that assesses search results to flag websites that host hoaxes, conspiracy theories, and "low-quality" content. Here, the "wetware" does the cultural work that the algorithm is unable to process.

However, and despite significant efforts in this area, there are ample reports that this content is still finding its way into news feeds and algorithmic recommendations. Platform companies have to deal with changing strategies on the part of content producers, trolls, and amplifiers as well as a growing search engine optimization industry.[87] Because social media and search engines still rely on data-driven signals to determine importance or relevance, false content continues to not only exist on these platforms, but also to trend. Search results are particularly malleable during times of crisis or confusion, such as following the Las Vegas shooting in October 2017 or the mass shooting in Texas, after which Google prioritized disinformation for a short period, before replacing it with fact-checked sources.[88]

De-prioritizing content and removing accounts also requires that platform companies make and stick to specific guidelines about acceptable content, which requires balancing the viewpoints and needs of many conflicting ideological communities. These platforms become the territories where these battles play out; where not only do content producers fight for attention, but everyday consumers do too. As corporations work to establish boundaries and guidelines for content, we must remain mindful that one universal standard for regulating all content online is shortsighted.[89] For instance, Twitter recently blocked search results for terms like "bisexual" after the keyword became associated with pornographic material by moderation processes.[90] This has sparked criticism from communication scholars, Safiya Noble and Sarah Roberts, who find the "commercial content moderation" guidelines used by these companies to be culturally insensitive and oblivious to harms they cause.[91] This can include the real pain of erasure as in the case of bisexuals, the hypersexualizing of ethnic groups and young women, or allowing for targeted harassment campaigns as in the case of Gamergate. These issues impact both automated content removal and human-led content moderation policies. The rules used to make these decisions have been panned for being both too opaque to be criticized and too vague to be applied consistently.[92]

Whether moderation is conducted by humans enforcing guidelines or by algorithms programmed by human values, discretion is an artifact of human decision-making which is reified by the choices platforms make to curate content. Users cannot flag everything in need of moderation, nor can moderators work consistently under the current tidal wave of user-generated content.[93] Moreover, Roberts' research illustrates that commercial content moderation has largely been outsourced by platform companies to independent contractors. The failure to recognize this new workforce has deleterious consequences not only for the industry, but also for moderators' psychological safety.[94] As such, it is not just the practices of content moderation that calls for a reckoning; the problem also requires concerted attention to the labor involved in ensuring policies and standards are applied consistently by a well-trained work force with access to healthcare and paid leave for psychological distress. Algorithms trained on gore, violence, and nudity can be a helpful buffer to filter out some graphic content. Alternatively, if the work of moderating content is to be contracted out, we are left with another critical question: What is the best organizational model that would ensure content meets the standards set by platform corporations, mitigates harm to moderators, and ensures a quality customer experience?

Whether moderation is conducted by humans enforcing guidelines or by algorithms programmed by human values, discretion is an artifact of human decision-making which is reified by the choices platforms make to curate content.

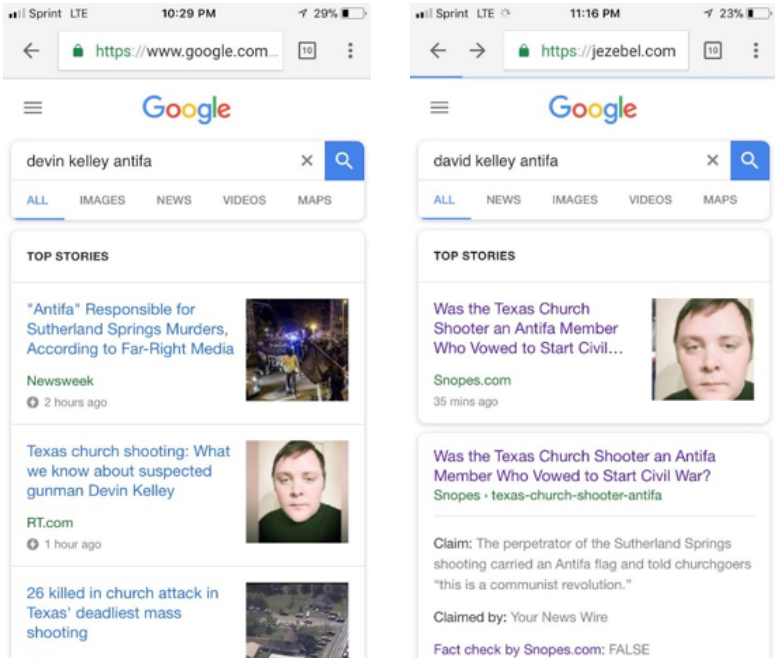


Fig. 6: Representative search results following Texas shooting.

STRATEGY 4: REGULATORY APPROACHES

Many of the disinformation efforts associated with "fake news" are tied to state-sponsored actions by the Russian government. Here, no definition of "fake news" can cover the enormous fissure in U.S. legal regimes pitting free speech values against invading international cyber-armies or domestic trolls. Government officials both within the United States and abroad have thus framed the problem of "fake news" as not only a problem for democracy, but as a tool of state-sponsored cyberwarfare. Acknowledging that market-based solutions cannot mitigate these threats, governments around the world are taking steps to address "fake news" and hate speech online through legislation, hearings, or establishing centers dedicated to the problem.

Though some countries, such as Germany, have recently passed legislation to fine social media companies and search engines for misleading content that engages in hate speech, the United States has not yet taken a regulatory approach towards moderating content on the internet that is not specifically illegal. This is because, within the United States, platform corporations have limited liability for content posted by users, due to a provision within the Communications Decency Act, known as Section 230, passed in 1996. This provision gives immunity and limited liability to platforms for content posted by users onto their networks.[95] Platform corporations or "interactive computer services" are allowed to self-regulate content due to a "Good Samaritan" provision within the same act, which gives platform owners protection should they voluntarily take action to "restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable" whether or not that content is constitutionally protected by the First Amendment. It is this law that gives platform corporations the power to moderate content according to their own standards. When signing up for services on a platform, everyone agrees to the same contract informing the user of their status as a customer and the "terms of service." Here, a platform's "terms of service" and community guidelines are the key mechanisms of enforcement, not the law.

Despite the legal protection for self-regulation and elaborate schemas for moderating commercial content, platform corporations *appear* to be very reluctant to moderate misleading content and hate speech.[96] However, researchers currently have no scale to measure how much content platforms remove, censor, or de-prioritize on any given day. As Nabiha Syed has argued, social media platforms in particular have embraced First Amendment theory not only in the governance of their content and content moderation policies, but as part of the mission of their companies.[97] Platform companies position themselves as an infrastructure for the marketplace of ideas, rather than editors or publishers making content decisions.[98] Though they have been reluctant over the last year to be the "arbiters of truth" when it

comes to determining what news content is legitimate or not, corporations like Facebook, Alphabet/Google, and Twitter seem to be more inclined to play this role as more misleading content is tied to state-backed propaganda efforts. In a series of hearings with members of Congress on October 31 and November 1, 2017, into the role of Russian disinformation efforts, representatives from these major platform corporations were questioned by senators on their efforts to limit the spread of misinformation and disinformation, specifically from state-backed sources. Twitter's acting general counsel, Sean Edgett, pledged they would develop a strategy to combat disinformation, stating "The abuse of our platform to attempt state-sponsored manipulation of elections is a new challenge for us, and one that we are determined to meet."[99]

However, without systemic oversight and auditing of platform companies' security practices, information warfare will surely intensify.

Because of Section 230 and the First Amendment, regulators within the United States currently have little recourse with which to limit the spread of misleading content. Even in instances where the object of regulation is state-sponsored information warfare campaigns, regulators may find it difficult to distinguish these efforts from those of American citizens, potentially leading to extended debates about First Amendment protections.[100] Senators within Congress are proposing limited regulations, such as the Honest Ads Act, to target certain types of misleading content within the realm of political advertisements–an area that has been regulated in the media industry before over print, cable, and broadcast.[101] Though senators made vague threats of regulation during the Senate Intelligence hearings, because of First Amendment protections, it is likely that platform corporations will have greater recourse to regulate content voluntarily through the Good Samaritan provision of Section 230, than Congress would have to regulate content over platforms. This may be, however, an untenable solution, leaving U.S. national security in the hands of private companies who are not incentivized to proactively look for problems playing out on their networks. However, without systemic oversight and auditing of platform companies' security practices, information warfare will surely intensify.

Lawmakers abroad are also taking actions against platforms. In June, the German parliament approved a bill (the NetzDG law) to limit the spread of hate speech and criminal material over social media, requiring social media platforms to remove these types of posts within 24 hours after receiving a notification or complaint or to block the offending content within seven days.[102] The law has been reported to be primarily directed at "fake news" that uses inflammatory and defamatory language directed at minorities, which was reported to have spread throughout Germany in the wake of the refugee crisis.[103] Social media companies face fines of up to 50

million euro if they "persistently fail to remove illegal content." The NetzDG law also requires social media companies, like Facebook and Twitter, and platforms, like Google, to remove "unlawful content" such as speech that includes a "public incitement to crime," defamation, "treasonous forgery," or depictions of violence."[104] However, it has also been widely criticized because it may put *more* power to censor content with platforms, as Germany is expecting platforms to largely adjudicate whether content should remain online. After the Unite the Right rally in Charlottesville, where activist Heather Heyer was murdered in an automobile attack on counter-protesters, a number of extremist far-right groups were "no platformed" by many internet sites and services.[105] The enforcement of terms of service against hate speech and symbols continues to be a contentious issue for platform companies. Twitter is currently leading this charge by taking steps to reduce the prominence of white supremacists and serial harassers on their site.[106]

How platform companies implement regulations regarding speech globally will determine who will govern the standards for speech in the future.

The lack of U.S. law places these decisions squarely under the jurisdiction of platforms as well, with little to no oversight. The lack of law here could mean that the German law becomes the default standard that is adopted throughout the world. Though platforms could use geo-blocking software to ensure these standards are *only* being adopted in combination with human moderation teams specifically to review content in Germany, the Good Samaritan provision within the United States gives platform companies the leeway to moderate content anywhere under the same guidelines they are using to moderate content in Germany. In the past, platform companies have adopted standards developed in courts abroad to inform the design of information systems in the United States.[107] And Mark Zuckerberg himself recently spoke of Facebook as encapsulating a "global community" that must combat the forces of "authoritarianism, isolationism and nationalism."[108] But while Facebook and other platform companies may be excited to frame themselves as defenders of society, this puts that society in a precarious position, with the same companies functioning as both advocates and owners. In any case, how platform companies implement regulations regarding speech globally will determine *who* will govern the standards for speech in the future.

CONCLUSION: MODERATING "FAKE NEWS" WILL IMPACT MORE THAN JUST NEWS

The use of the phase "fake news" by social media scholars, journalists, and industry members refers to content that is believed to be propaganda, false, or spread with negative intentions. Currently, these scholars, advocates, and members of industry are working diligently to differentiate such "fake news" content from real news by determining the intent of the speaker, analyzing the range of types of content, or identifying feature-based characteristics that can be used to identify untrustworthy content.

For right-wing media makers, "fake news" is increasingly used to denote content that should be considered illegitimate because of its association with left or liberal political leanings. In some cases, President Trump has used explainable mistakes in reporting to delegitimize entire news outlets; *CNN* has been a particularly prominent target.[109] Other times, "fake news" is used as an extension of well-worn critiques that the mainstream media feeds audiences false and untrustworthy narratives.[110] As a result, this criticism serves as a justification for an alternative media network of hyper-partisan and conspiracy laden news sources that often spread disinformation and hoaxes.[111] Efforts to combat "fake news" over social media will be viewed by these groups as a partisan effort, directed at *their alternative* news outlets rather than others.[112] Yet, when taking into consideration typology approaches to defining "fake news," a story like "Pizzagate" delivers a broadside; it is this alternative media network, alongside Russian actors and anonymous trolls, that often spreads "fake news."

Still, when looking at the range of definitions and interventions targeting "fake news," we find much more than a simple partisan binary. Even groups that agree on what "fake news" is disagree on its importance or how to stop it. "Fake news" as a keyword in public discourse offers a proxy view of global techno-politics, where different values guide beliefs about what content should be moderated, who should be responsible, how moderation should be applied to different groups, and what kinds of mental and physical harms are tolerated. "Fake news" is more than a widening of partisanship and the misleading use of social media to spread disinformation; it's about the social sharing of trust, credibility, and evidence in the making of an informed citizenry.

The way through this dead reckoning remains foggy, at best. The problems associated with "fake news" appear moored to platform corporations' business models, where

monetary incentives to manipulate information systems subsist alongside political and cyberwarfare campaigns. Interestingly, subscription-based news media received a bump after Trump's election.[113] However, it was not enough to significantly increase circulation rates or reel in new advertising revenue.[114] While establishing standards for news media is desperately needed across the platform companies, it is vital to think about how defining "fake news" will asymmetrically impact small-market press and independent journalists, who cannot afford the advertising necessary to grow audiences or rely on organic reach. Strategies to intervene on "fake news" should consider rewarding good news as much as they do punishing the bad. New strategies could include incentives and rewards for improving the quality of information online, while also seeking out and penalizing concerted disinformation operations.

Moreover, designing interventions should employ the broadest possible coalition across some high-tension zones, where the concerns about growing networks of hate movements and enhanced online surveillance through targeted data harvesting – brought to the fore by groups like Color of Change, Center for Media Justice, Center for Democracy and Technology, Southern Poverty Law Center, and Free Press – are taken into account. New interventions must involve protections from digital harms for news consumers, content moderators, and journalists. Taking a multi-stakeholder approach will not rid platform companies of user-generated content rife with racism, homophobia, misogyny, or harassment, but it will make it harder for groups espousing these reprehensible beliefs to find shelter by labeling themselves news organizations.

With "fake news," the risk is not necessarily that it will overtake real news, but that democracy itself might drown in information. Those unable to assess and critique online content for its veracity and journalistic integrity will run aground on hyper-partisan media sources that are trusted amongst members of their community.[115] Without employing standards for what counts as news, societies lose the basic materials for democratic decision making. If we are to break out of the ironclad echo chambers that pattern online information, interventions must begin with acknowledging a free press as an anchor for democratic societies, while also determining how online media is manipulated for different ends; and finally, charting a course for what can be done to ensure accountability across the entire news industry through cycles of content production, delivery, and consumption. To be sure, moving towards the offing will require all hands on deck to map new routes, create new technologies, and enforce new standards; which means fundamental organizational changes to platform corporations who seek to build a global community. What is at stake today in fighting "fake news" will not only decide whose voices matter and whose voices are worth amplifying, but also who gets to build communication technology, who gets to scale it, and for whom it is most useful.

ACKNOWLEDGMENTS

This report draws from the research and insights of the Media Manipulation Initiative at Data & Society. Thank you to Becca Lewis, Caroline Jack, Matt Goerzen, Patrick Davison, and Francesca Tripodi for your contributions. Thank you to Patrick Davison, Alice Marwick, and danah boyd for their extensive edits and feedback, and to external contributors, Monica Bulger, Mark Ackerman, C.W. Anderson, and Philip M. Napoli, for their insights and guidance.

Illustrations by Jim Cooke
Design by Jeff Ytell

ENDNOTES

1. Craig Silverman, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook," *BuzzFeed News*, November 16, 2016, https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm_term=.nr5Wrr1Grm#.oaLzxxB7x9.

2. Olivia Solon, "Facebook's failure: Did fake news and polarized politics get Trump elected?" *The Guardian*, November 10, 2016, https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories.

3. Craig Silverman and Lawrence Alexander, "How Teens in the Balkans are Duping Trump Supporters with Fake News," *BuzzFeed News*, November 3, 2016, https://www.buzzfeed.com/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo?utm_term=.ng2pWWq7W4#.haR3VVMAVR.

4. Olivia Solon, "Facebook's Fake News: Mark Zuckerberg rejects 'crazy idea' that it swayed voters," *The Guardian*, November 10, 2016, https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-us-election-mark-zuckerberg-donald-trump.

5. It is unclear whether Trump knew of the other connotations of the word before he decided to use it against *CNN* and *The New York Times*. In an interview with the Christian network Trinity Broadcasting, on October 8, 2017, Trump claimed he had "come up with" the term "fake" in referring to news. Read more at *New York Daily News* http://www.nydailynews.com/news/politics/trump-created-word-fake-defends-puerto-rico-towels-article-1.3549242?utm_content=buffer89246&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer.

6. Bret Stephens, "The President Versus 'Fake News," Again." *The New York Times*, June 29, 2017, https://www.nytimes.com/2017/06/29/opinion/trump-cnn-fake-news-russia.html.

7. Paul Farhi, "Committee Rejects Breitbart Application for Congressional Press Credentials," *The Washington Post*, April 25, 2017, https://www.washingtonpost.com/lifestyle/style/committee-rejects-breitbart-application-for-congressional-press-credentials/2017/04/25/70c80992-29c8-11e7-b605-33413c691853_story.html?utm_term=.25efb3f1f449.

8. Craig Silverman, "BuzzFeed News: Election content engagement," November 16, 2016, https://docs.google.com/spreadsheets/d/1ysnzawW6pDGBEqbXqeYuzWa7Rx2mQUip6CXUUUk-4jIk/edit#gid=1756764129.

9. Claire Wardle, "Fake News. It's Complicated," *First Draft News*, February 16, 2017, https://firstdraftnews.com/fake-news-complicated/.

10. For more on the effects of profitability and advertising on search results, see Noble, Safiya. "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." *InVisible Culture: An Electronic Journal for Visual Culture*, no. 19 (October 30, 2013). https://urresearch.rochester.edu/institutionalPublicationPublicView.action;jsessionid=3981AF546CDC3311879B89ED791B823B?institutionalItemId=27584&versionNumber=1. See also: Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press, 2018.

11. Barney Glaser and Anselm Straus, 2017. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Publishing Company.

12. Amelia Acker and Brian Beaton, "How Do You Turn a Mobile Device into a Political Tool?," University of Hawai'i, 2017. https://doi.org/10.24251/HICSS.2017.281.

13. Steve Coll, "Donald Trump's 'Fake News' Tactics," *The New Yorker*, December 11, 2017, https://www.newyorker.com/magazine/2017/12/11/donald-trumps-fake-news-tactics.

14. Claire Wardle and Hossein Derakshan, "Information Disorder: Toward an interdisciplinary framework for research and policy making," *Council of Europe*, September 27, 2017, https://firstdraftnews.com/wp-content/uploads/2017/10/Information_Disorder_FirstDraft-CoE_2018.pdf?x40896.

15. Prager University, "What is Fake News," *PragerU.com*, June 29, 2017. https://www.prageru.com/courses/political-science/what-fake-news.

16. Nicole Hemmer, *Messengers of the Right: Conservative Media and the Transformation of American Politics* (Philadelphia: University of Pennsylvania Press, 2016).

17. Michael M. Grynbaum, "A Costly Retraction for CNN and an Opening for Trump," *The New York Times*, June 27, 2017, https://www.nytimes.com/2017/06/27/business/media/cnn-retracted-story-on-trump.html.

18. Yochai Benkler, Robert Faris, Hal Roberts, and Ethan Zuckerman, "Study: Breitbart-led right-wing media ecosystem altered broader media agenda," *Columbia Journalism Review*, March 3, 2017, https://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php.

19. Nicholas Confessore, and Daisuke Wakabayashi, "How America Harvested American Rage to Reshape U.S. Politics." *The New York Times*. October 9, 2017. https://www.nytimes.com/2017/10/09/technology/russia-election-facebook-ads-rage.html?_r=0.

20. Samanth Subramanian. "Inside the Macedonian Fake-News Complex." *Wired.com*. February 15, 2017. https://www.wired.com/2017/02/veles-macedonia-fake-news/.

21. Mike Cernovich. "Busted! Fake news blog the Daily Beast lied about Mike Cernovich." *Medium.com*, December 13, 2017. https://medium.com/@Cernovich/busted-fake-news-blog-the-daily-beast-lied-about-mike-cernovich-89c21b374593.

22. There were several attempts to debunk Pizzagate, see below: Gregor Aisch, "Dissecting the #PizzaGate Conspiracy Theories." *The New York Times*, December 10, 2016, sec. Business Day. https://www.nytimes.com/interactive/2016/12/10/business/media/pizzagate.html; James Doubek, "Conspiracy Theorist Alex Jones Apologizes For Promoting 'Pizzagate.'" NPR.org. Accessed December 30, 2017. https://www.npr.org/sections/thetwo-way/2017/03/26/521545788/conspiracy-theorist-alex-jones-apologizes-for-promoting-pizzagate; Kim LaCapria, "FALSE: Comet Ping Pong Pizzeria Home to Child Abuse Ring Led by Hillary Clinton." Snopes.com, November 21, 2016. https://www.snopes.com/pizzagate-conspiracy/.

23. Carlett Spike and Pete Vernon, "'It Was Super Graphic': Reporters Reveal Stories of Online Harassment." Columbia Journalism Review. Accessed December 31, 2017. https://www.cjr.org/covering_trump/journalists-harassment-trump.php; Terry Gross, "Harassed On Twitter: 'People Need To Know The Reality Of What It's Like Out There.'" NPR.org. Accessed December 31, 2017. https://www.npr.org/2016/10/26/499440089/harassed-on-twitter-people-need-to-know-the-reality-of-what-its-like-out-there.

24. For an example see Tim Wu, "Is the First Amendment Obsolete," *Knight First Amendment Institute*, September, 2017, https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete.

25. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Lu, "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations Newsletter* (2017): arXiv:1708.01967.

26. Claire Wardle, "Fake News. It's Complicated," *First Draft News*, February 16, 2017, https://firstdraftnews.com/fake-news-complicated/.

27. Mark Verstraete, Derek E. Bambauer, and Jane R. Bambauer, "Identifying and Countering Fake News," *Arizona Legal Studies*, 17-75 (2017).

28. B.S. Detector. http://bsdetector.tech/.

29. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Lu, "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations Newsletter* (2017): arXiv:1708.01967.

ENDNOTES

30. Caroline Jack, "Lexicon of Lies: Terms for Problematic Information," *Data & Society Research* Institute, 2017, https://datasociety.net/pubs/oh/DataAndSociety_LexiconofLies.pdf.

31. Fkues@poynter.org, "What's the legal definition of "fake news?" One newspaper publisher may sue to find out." *Poynter.org*, March 29, 2017, https://www.poynter.org/news/whats-legal-definition-fake-news-one-newspaper-publisher-might-sue-find-out.

32. David O. Klein and Joshua R. Weller, "Fake News: A Legal Perspective," *Journal of Internet Law*, 20, 1 (2017).

33. Edson C. Tandoc Jr., Zheng Wei Lim and Richard Ling, "Defining 'Fake News,' *Digital Journalism* (2017): DOI: 10.1080/21670811.2017.1360143.

34. Emma Grey Ellis, "Can't Take a Joke? That's Just Poe's Law, 2017's Most Important Internet Phenomenon," *Wired Magazine*, June 5, 2017, https://www.wired.com/2017/06/poes-law-troll-cultures-central-rule/.

35. Mark Verstraete, Derek E. Bambauer, and Jane R. Bambauer, "Identifying and Countering Fake News," *Arizona Legal Studies*, 17-75 (2017).

36. Claire Wardle, "Fake News. It's Complicated," *First Draft News*, February 16, 2017, https://firstdraftnews.com/fake-news-complicated/.

37. Cale Guthrie Weissman, "Watch Out For This Fake News Website Masquerading as The New York Times," *Business Insider*, June 29, 2015, http://www.businessinsider.com/nytimescomco-posts-fake-news-articles-pretending-to-be-the-new-york-times-2015-6.

38. Todd Spangler, "NBCU Sends Cease-and-Desist Notice to Fake News Site," *Variety*, October 29, 2015, http://variety.com/2015/digital/news/nbcu-fake-news-breaking-bad-season-6-legal-cease-desist-1201629813/.

39. Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. "Evaluating Information: The Cornerstone of Civic Online Reasoning," *Stanford Digital Repository* (2016): https://purl.stanford.edu/fv751yt5934.

40. *The Valley Report*, "Disclaimer," https://thevalleyreport.com/disclaimer/. Archived on archive.is on November 16, 2017 at https://archive.fo/8WzeU.

41. Snopes, "Were Police Told to 'Stand Down' Amid Violence in Charlottesville?" *Snopes*, https://www.snopes.com/were-police-told-stand-down-charlottesville/.

42. Kai Shu, Amy Silva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective," *arXiv* (2017): 1708.01967v3.

43. Meedan, "Building Technical Standards for Credibility," *MisinfoCon.com*, March 2, 2017, https://misinfocon.com/building-technical-standards-for-credibility-59ef9ee4ab73.

44. Interview with Cameron Hickey of PBS NewsTracker by Robyn Caplan from Data & Society Research Institute, on Friday October 13th over *Slack*.

45. Casey Newton,"Facebook is patenting a tool that could help automate removal of fake news," *The Verge*, December 7, 2016, https://www.theverge.com/2016/12/7/13868650/facebook-fake-news-patent-tool-machine-learning-content.

46. Facebook, Inc, Erez Laks, Adam Stopeck, Adi Masad, Israel Nir, "Systems and Methods to Identify Objectionable Content," *United States Patent Application*, June 1, 2015, http://pdfaiw.uspto.gov/.aiw?PageNum=0&docid=20160350675&IDKey=B0738725A-3CA&HomeUrl=http%3A%2F%2Fappft.uspto.gov%2Fnetacgi%2Fnph-Parser%3FSect1%3DPTO1%2526Sect2%3DHITOFF%2526d%3DPG01%2526p%3D1%2526u%3D%2Fnetahtml%2FPTO%2Fsrchnum.html%2526r%3D1%2526f%3DG%2526l%3D50%2526s1%3D20160350675.PGNR.%2526OS%3D%2526RS%3D.

47. Sam Levin, Julia Carrier Wong, and Luke Harding, "Facebook Backs Down from 'Napalm Girl's' Censorship and Reinstates Photo," *The Guardian*, September 9, 2016, https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo.

48. Julia Angwin and Hannes Grasseger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children," *ProPublica*, June 28, 2017, https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

49. Global Voices, "Style Guide." *Global Voices Community Blog*. https://community.globalvoices.org/guide/editorial-guides/style-guide/#Anonymous_Authors.

50. Andrew Liptak, "Mark Zuckerberg Warns About Facebook Becoming 'Arbiters of Truth.'" *The Verge*, November 13, 2016, https://www.theverge.com/2016/11/13/13613566/mark-zuckerberg-facebook-misinformation-hoax-media.

51. Robyn Caplan, "How Do You Deal With a problem Like 'Fake News?'" *Points // Data & Society*, January 5, 2017, https://points.datasociety.net/how-do-you-deal-with-a-problem-like-fake-news-80f9987988a9.

52. Edson C. Tandoc Jr., Zheng Wei Lim & Richard Ling, "Defining 'Fake News,'" *Digital Journalism* (2017): DOI: 10.1080/21670811.2017.1360143.

53. Phil M. Napoli and Robyn Caplan, "Why Media Companies Insist They're Not Media Companies, Why They're Wrong, and Why It Matters," *First Monday*, 22, 5, (2017): http://firstmonday.org/ojs/index.php/fm/article/view/7051.

54. David Kravets, "DMCA 'Safe Harbor' up in the Air for Online Sites That Use Moderators." *Ars Technica*, April 10, 2017. https://arstechnica.com/tech-policy/2017/04/dmca-safe-harbor-up-in-the-air-for-online-sites-that-use-moderators/.

55. Michael Nunez, "Former Facebook Workers: We Routinely Suppressed Conservative News." *Gizmodo*, September 5, 2016, https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006.

56. Lucas Graves, *Deciding What's True* (New York: Columbia University Press, 2016).

57. *Fake News Watch*. "RealNewsRightNow.com," http://fakenewswatch.com/realnewsright-now-com. (accessed September 2017).

58. Jay Hathaway, "The Antiviral Guide to the Worst Hoaxers and Liars on Facebook," *AntiViral. Gawker.com*, March 3, 2015, https://digg.com/source/antiviral.gawker.com; see also Caitlin Dewey, "What was fake on the Internet this week: Belle Gibson, ISIS in Mexico and insane weather news," The Washington Post, April 24, 2015, https://www.washingtonpost.com/news/the-intersect/wp/2015/04/24/what-was-fake-on-the-internet-this-week-belle-gibson-isis-in-mexico-and-insane-weather-news/.

59. *First Draft News*, "About," https://firstdraftnews.com/about/.

60. Robyn Caplan, "How Do You Deal with a Problem Like 'Fake News?'" *Data & Society*, January 5, 2017, https://points.datasociety.net/how-do-you-deal-with-a-problem-like-fake-news-80f9987988a9. See Tessa Lyons, "News Feed FYI: Replacing Disputed Flags with Related Articles," Facebook, December 20, 2017, https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/.

61. Natasha Lomas, "Facebook Takes Its Fake News Fight to Germany," *TechCrunch*, January 16, 2017, https://techcrunch.com/2017/01/16/facebook-takes-its-fake-news-fight-to-germany/.

62. Markkula Center for Applied Ethics. "The Founding of the Trust Project," *Santa Clara University*, 2017, https://www.scu.edu/ethics/focus-areas/journalism-ethics/resources/the-founding-of-the-trust-project/.

63. Sarah Perez, "Facebook, Google and others join The Trust Project, an effort to increase transparency around online news," *TechCrunch*, 2017, November 16, https://techcrunch.com/2017/11/16/facebook-google-and-others-join-the-trust-project-an-effort-to-increase-transparency-around-online-news/.

64. Tyler Durden, "Exposing the 9 Fakest Fake-News Checkers." *ZeroHedge.com*, February 20, 2017, http://www.zerohedge.com/news/2017-02-20/exposing-9-fakest-fake-news-checkers.

65. Gordon Pennycook, Tyrone D. Cannon, and David G. Rand, "Implausability and Illusory Truth: Prior Exposure Increases Perceived Accuracy of Fake News but Has No Effect on Entirely Implausible Statements," *SSRN* (2017), https://ssrn.com/abstract=2958246.

66. Gordon Pennycook and David G. Rand, "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings," *SSRN* (2017), https://ssrn.com/abstract=3035384.

67. Gian Fulgoni, "Fraud in Digital Advertising: A Multibillion-Dollar Black Hole," *Journal of Advertising Research* (2016): http://www.journalofadvertisingresearch.com/content/56/2/122.

68. Gordon Pennycook and David G. Rand, "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings," Available at *SSRN* (2017), https://ssrn.com/abstract=3035384.

69. Suzanne Vranica, "Advertisers Try to Avoid the Dark Side, From Fake News to Extremist Videos," *The Wall Street Journal*, June 18, 2017, https://www.wsj.com/articles/advertisers-try-to-avoid-the-webs-dark-side-from-fake-news-to-extremist-videos-1497778201.

70. Jack Nicas, "Google's YouTube Has Continued Showing Brands' Ads with Racist and Other Objectionable Content," *The Wall Street Journal*, March 24 2017, https://www.wsj.com/articles/googles-youtube-has-continued-showing-brands-ads-with-racist-and-other-objectionable-videos-1490380551.

71. Julia Love and Kristina Cooke, "Google and Facebook Are Cracking Down to Prevent Their Ads Appearing on Fake News Sites," *Reuters/Business Insider*, November 15, 2016, http://www.businessinsider.com/google-facebook-crack-down-adverts-appearing-fake-news-sites-us-election-trump-2016-11.

72. YouTube, "Advertiser-Friendly Content Guidelines," *YouTube*, 2016, https://support.google.com/youtube/answer/6162278?hl=enhttps://support.google.com/youtube/answer/6162278?hl=en.

73. Lizzie Plaugic, "YouTube Creators Are Frustrated That a Bot Keeps Demonetizing Their Videos," *The Verge*, November 14, 2017, https://www.theverge.com/2017/11/14/16648348/youtube-demonetizing-iphone-x-videos.

74. Disclosure: Data & Society is part of the News Integrity Initiative.

75. News Corp. "Storyful and Moat Launch Initiative to Combat Fake News," *NewsCorp.com*, 2017. https://newscorp.com/2017/05/02/storyful-and-moat-launch-initiative-to-combat-fake-news/.

76. Janna Anderson and Lee Rainie, "The Future of Truth and Misinformation Online," *Pew Research Center*, October 19, 2017. http://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/?utm_source=adaptivemailer&utm_medium=email&utm_campaign=10/19/2017%20foi%20misinformation&org=982&lvl=100&ite=1868&lea=395500&ctr=0&par=1&trk=.

77. Jen Weedon, William Nuland, and Alex Stamos, "Information Operations and Facebook," *Facebook Newsroom*, April 27, 2017, https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf.

78. Eric Auchard and Joseph Menn, "Facebook Cracks Down on 30,000 Fake Accounts in France," *Reuters*, April 13, 2017, https://www.reuters.com/article/us-france-security-facebook/facebook-cracks-down-on-30000-fake-accounts-in-france-idUSKBN17F25G.

79. "How and When Are My Tweets Not Seen by Everyone?" *Twitter Support*, https://support.twitter.com/articles/20175217.

80. Tess Townsend, "Google Has Banned 200 Publishers Since it Passed a New Policy Against Fake News," *Recode*, January 25, 2017, https://www.recode.net/2017/1/25/14375750/google-adsense-advertisers-publishers-fake-news.

81. Norbert Wiener, *The Human Use of Human Beings: Cybernetics and Society*, Da Capo Press, 1988. Print.

82. Adam Mosseri, "Working to Stop Misinformation and False News," *Facebook Newsroom*, April 6, 2017, https://newsroom.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/.

83. Erich Owens and Udi Weinsberg, "News Feed FYI: Showing Fewer Hoaxes," *Facebook Newsroom*, January 20, 2015, https://newsroom.fb.com/news/2015/01/news-feed-fyi-showing-fewer-hoaxes/.

84. Jiun-Ren Lin and Shengbo Guo, "News Feed FYI: Reducing Links to Low-Quality Web Page Experiences." *Facebook Newsroom*, May 10, 2017, https://newsroom.fb.com/news/2017/05/reducing-links-to-low-quality-web-page-experiences/.

85. Christopher Hein, "Facebook is Now Reducing the Reach of Users Who Routinely Share Fake News, Clickbait and Spam," *Adweek*, June 30, 2017, http://www.adweek.com/digital/facebook-is-reducing-the-reach-of-users-who-routinely-share-fake-news-clickbait-and-spam/.

86. Mark Bergen, "Google Rewrites Its Powerful Search Rankings to Bury Fake News," *Bloomberg*, April 25, 2017, https://www.bloomberg.com/news/articles/2017-04-25/google-rewrites-its-powerful-search-rankings-to-bury-fake-news.

87. Sara Fischer, "Fake News Takes the World," *Axios*, October 3, 2017, https://www.axios.com/fake-news-takes-the-world-2492360210.html.

88. Janna Anderson and Lee Rainie, "The Future of Truth and Misinformation Online," *Pew Research Center*, October 19, 2017, http://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/?utm_source=adaptivemailer&utm_medium=email&utm_campaign=10/19/2017%20foi%20misinformation&org=982&lvl=100&ite=1868&lea=395500&ctr=0&par=1&trk= .

89. Julia Angwin and Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men from Hate Speech but not Black Children," *ProPublica*, June 28, 2017, https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

90. Charlie Warzel, "Big Tech Cannot Stop Shooting Itself in the Foot," *BuzzFeed News*, November 17, 2017, https://www.buzzfeed.com/charliewarzel/big-tech-cannot-stop-shooting-itself-in-the-foot?utm_term=.nkokvvBYvE#.ed9gJJQ4Jq.

91. For more about commercial content moderation, see: Sarah T. Roberts, "Content Moderation."

In *Encyclopedia of Big Data*, edited by Laurie A. Schintler and Connie L. McNeely, 1–4. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-32001-4_44-1. For more information on the harms caused by algorithmic filtering, see: Safiya Umoja Noble, Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press, 2018.

92. Nick Statt, "Twitter Releases Its Calendar of Upcoming Measures to Combat Harassment and Abuse," *The Verge*, October 19, 2017, https://www.theverge.com/2017/10/19/16505954/twitter-harassment-abuse-calendar-schedule-fixes-updates.

93. Sarah T. Roberts, "Content Moderation." In *Encyclopedia of Big Data*, edited by Laurie A. Schintler and Connie L. McNeely, 1–4. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-32001-4_44-1.

94. For more about the psychological costs associated with content moderation by humans, see: Sam Levin, "Moderators Who Had to View Child Abuse Content Sue Microsoft, Claiming PTSD." *The Guardian*, January 12, 2017, sec. Technology. http://www.theguardian.com/technology/2017/jan/11/microsoft-employees-child-abuse-lawsuit-ptsd.

95. Communications Decency Act, 47 U.S.C. §230

96. Sarah T. Roberts, "Commercial Content Moderation: Digital Laborers' Dirty Work," *Media Studies Publications*, 12 (2016): http://ir.lib.uwo.ca/commpub/12.

97. Nabiha Syed, "Real Talk about Fake News: Towards a better theory for platform governance," *The Yale Law Journal*, 127 (2017): https://www.yalelawjournal.org/forum/real-talk-about-fake-news.

98. Unless legally required, such as is the case with child pornography, or due to business imperatives, such as nudity/pornography, or when asked to remove it by state actors, such as with terrorist content.

99. Cecilia Kang, Nicholas Fandos, and Mike Isaac, "Tech Executives are Contrite About Election Meddling, but Make Few Promises on Capitol Hill," *The New York Times*, October 31, 2017, https://www.nytimes.com/2017/10/31/us/politics/facebook-twitter-google-hearings-congress.html.

100. Tim Wu, "Is the First Amendment Obsolete?" *Knight First Amendment Institute*, September, 2017. https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete.

101. Laura Hazard Owen, "The Honest Ads Act Would Force Internet Companies to Change Their Disclosure Practices by January 2018," *Nieman Lab*, October 20, 2017. http://www.niemanlab.org/2017/10/the-honest-ads-act-would-force-internet-companies-to-change-their-disclosure-practices-by-january-2018/.

102. The Guardian, "Germany Approves Plans to Fine Social Media Firms Up to 50m," *The Guardian*, June 30, 2017, https://www.theguardian.com/media/2017/jun/30/germany-approves-plans-to-fine-social-media-firms-up-to-50m.

103. Mark Scott, "Ahead of Election, Germany Seeks Fake News Antidote," *Politico*, August 31, 2017, https://www.politico.eu/article/germany-election-campaign-fake-news-ange-

la-merkel-trump-digital-misinformation/.

104. Center for Democracy and Technology, "Overview of the NetzDG Network Enforcement Law," *CDT.org*. July 17, 2017. https://cdt.org/insight/overview-of-the-netzdg-network-enforcement-law/.

105. See examples here: Keith Collins, "A running list of websites and apps that have banned, blocked, deleted, and otherwise dropped white supremacists." *Quartz*. August 16, 2017. https://qz.com/1055141/what-websites-and-apps-have-banned-neo-nazis-and-white-supremacists/.

106. Bob Moser, "How Twitter's Alt-Right Purge Fell Short." *Rolling Stone*. Accessed December 31, 2017. https://www.rollingstone.com/politics/news/how-twitters-alt-right-purge-fell-short-w514444.

107. "Mark Zuckerberg's Commencement Address at Harvard." *Harvard Gazette*, May 25, 2017. https://news.harvard.edu/gazette/story/2017/05/mark-zuckerbergs-speech-as-written-for-harvards-class-of-2017.

108. "Mark Zuckerberg's Commencement Address at Harvard." *Harvard Gazette*, May 25, 2017. https://news.harvard.edu/gazette/story/2017/05/mark-zuckerbergs-speech-as-written-for-harvards-class-of-2017.

109. Michael M. Grynbaum, "A Costly Retraction for CNN and an Opening for Trump," *The New York Times*, June 28, 2017, https://www.nytimes.com/2017/06/27/business/media/cnn-retracted-story-on-trump.html?_r=0.

110. PragerU, "What is Fake News," *PragerU.com*, June 29, 2017. https://www.prageru.com/courses/political-science/what-fake-news.

111. Amanda Robb, "Pizzagate: Anatomy of a Fake News Scandal." *Rolling Stone*. Accessed December 30, 2017. https://www.rollingstone.com/politics/news/pizzagate-anatomy-of-a-fake-news-scandal-w511904.

112. For an example of this issue, see the Gizmodo article by Michael Nunez, "Former Facebook Workers: We Routinely Suppressed Conservative News," *Gizmodo.com*, May 9, 2016, https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006. See also Yochai Benkler, Robert Faris, Hal Roberts, and Ethan Zuckerman, "Study: Breitbart-led right-wing media ecosystem altered broader media agenda," *Columbia Journalism Review*, March 3, 2017, https://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php.

113. Ken Doctor, "Trump Bump Grows into Subscription Surge – and Not Just for the New York Times." *TheStreet*, March 3, 2017. https://www.thestreet.com/story/14024114/1/trump-bump-grows-into-subscription-surge.html.

114. Michael Barthel, "Despite Subscription Surges for Largest U.S. Newspapers, Circulation and Revenue Fall for Industry Overall." *Pew Research Center* (blog), June 1, 2017. http://www.pewresearch.org/fact-tank/2017/06/01/circulation-and-revenue-fall-for-newspaper-industry/.

115. Yochai Benkler, Robert Faris, Hal Roberts, and Ethan Zuckerman, "Study: Breitbart-led right-wing media ecosystem altered broader media agenda," *Columbia Journalism Review*, March 3, 2017, https://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php