Literature review

The sheer volume of work on social media analysis involving topics such as extremes/extremism, mis/disinformation etc. and what might be done about it, makes it difficult to provide a concise yet fair review, while its recency makes it hard to identify seminal works. Hence we present here a literature review of the broad topics. It is of course also incomplete, despite running to many pages and references. It is also nuanced by opinion. We hope it may serve as a more general resource for researchers.

These sections were primarily written by our research assistants who were paid 100% by our research grants for this explicit purpose. Their names are available on request and will appear as an explicit acknowledgment to accompany any final published paper.

Sections:

1. Misinformation: Conceptualization and Trends

2. The Diffusion of Misinformation Online

3. Foreign Influence and Online Misinformation

4. Vulnerability to Misinformation in the U.S.

5. Models for content moderation

6. Strategies for content moderation and their effect on malicious content

7. How Individuals and Groups React to and Evade Content Moderation and How Platforms Respond

8. Contemporary Advancements in Online Content Moderation

9. Differences between Platforms

1. Misinformation: Conceptualization and Trends

Prepared primarily by SS

This section addresses two broad questions regarding the state of the literature on online misinformation. First, how is (mis)information—and its various subtypes—conceptualized both in theory and in practice. Second, what is all of the talk of society having entered a "post-truth" era really about?

The goal of this section is to review scholarship on misinformation that is situated at a fairly high level of abstraction (*what is "misinformation" really?*), but also to add in some empirical—but still grand theoretical—grounding (*what aspects of modern society make misinformation a hot issue?*).

1.1 **Conceptualizing Misinformation**

The concept of "misinformation" is an evolving one, both in its definitions as well as the normative connotations attached to those definitions. Scholarly interest in misinformation predates the digital age. One could find familiar intellectual roots even as far back as Ancient Greek fears that demagogues and rhetoricians might deceive or mislead the public (Plato 2000) as well as distinctions between informational appeals through *pathos*, *ethos*, and *logos* (Plato 1994; Aristotle 2007).

The modern focus on the spread and effects of unreliable or false information took off in the mid-20th century, though at the time such forms of misinformation were studied using a different vocabulary (one of "rumors" and "propaganda"). In the early 2000s, more contemporary terms like "fake news" began to enter the lexicon—though they were used to describe different forms of expression (more "satire" than malicious "deception") and did not yet have today's distinctly negative overtones. As we will see below, much of this older terminology continues to exert an influence on recent work.

Intellectual Shifts in the Literature

The modern study of *misinformation* can be traced back to research on *rumor* spread emerging at the end of WWII. During this period, scholarly interest in this subject (particularly in the field of psychology) was motivated by the concern that wartime rumors not only "impair public moral and confidence, but are in many circumstances the deliberate weapon of enemy *propaganda*" (Knapp 1944, 22). Rumormongering was viewed as an important misinformation problem of the age, one which individuals participated in out of a psychological desire to relieve their anxiety, justify their emotions, and extract meaning from their environments (Allport and Postman 1947).

Students of information sciences and technology continue to note that while misinformation is conceptually distinct from rumor (in that the latter *may or may not* be accurate), the 20th century rumor literature suggests promising avenues of inquiry for contemporary scholars by drawing attention to the motivational and emotional factors that drive public engagement with unreliable information (Chen and Sin 2013). Some recent work, while defining "rumor" and "misinformation" differently, still slides between these two terms (sometimes treating them interchangeably) when discussing methods for detecting and containing the spread of unverifiable information on Online Social Media (Qazvinian et al 2011). Other work tracks online rumor spread in a way that explicitly maintains conceptual distinctions between *rumor* and *misinformation*, operationalizing a rumor as online content that is "unverified" but which may ultimately turn out to be "true, or partly or entirely false" (Zubiaga et al 2015, 2016).

Older distinctions between different types of rumors have largely disappeared in recent work, but they may prove fruitful to consider even if in a limited—and necessarily altered—state to better fit new realities. One typology offered in post-WWII work on rumor spread (Knapp 1944), for instance, differentiated between three variants of rumors: (1) *the pipe-dream rumor*, which expresses the hopes of those who circulate it and is akin to wishful thinking (e.g., that Trump will be reinstated as President); (2) *the bogie or fear rumor*, which is derived from its propagators' fears and anxieties (e.g., that a migrant caravan will soon invade the U.S.); and (3) *the* *wedgedriving rumor*, which is motivated by aggression or hatred and has the effect of dividing groups and destroying loyalties (e.g., that Democrats are part of a devil worshiping cabal that eats children).

Recovering old insights on rumor diffusion may be helpful for new scholarship on online misinformation. This suggestion is in line with points made by some computer scientists who have begun to lament the blind spot on rumors in the current literature and, further, to highlight the "serious need for a better understanding of how fake news stories transform into rumors and to what extent these rumors can amplify beliefs and infiltrate other communities" (Ruths 2019).

And with this call for action, we have stumbled into a new term (*"fake news"*) which has recently come to acquire an explosive influence within the literature on online misinformation. References to fake news are now at the forefront of the misinformation research agenda, but this is a shift that occurred only within the past few years. Perhaps more notably, it occurred alongside a dramatic shift in the meaning of "fake news" from a relatively innocuous concept to something much more sinister.

In the early 2000s, the term "fake news" was most typically used to describe satirical TV-shows that had components of news parody (e.g., Saturday Night Live, The Daily Show, The Colbert Report). Such work defined fake news as a comedy sub-genre "driven by the presentation of political and journalistic satire... but [which virtually never] endorses a political candidate... [or] a given public policy issue" (Holbert 2005). Unsurprisingly, scholars writing about this conception of fake news most typically came from backgrounds in the study of media, journalism, and communication. Some contributors to this literature treated satirical programs like *The Daily Show* as providing much more "real" news content than others, viewing it as an innovative—albeit humorous and hyperbolic—experiment in a brand critical journalism that had begun to recede from conventional media sources (Baym 2005). Even so, debates about "fake news" until 2015 continued to revolve around satirical TV and radio programs as the objects of their inquiry (Day and Thompson 2012; Balmas 2014).

Fake news of this variety was not yet viewed as having inherently negative effects on social and political behavior. Some studies found that fake news (as a sub-genre of satire) could indeed produce negative cognitive reactions in its viewers (e.g., alienation, cynicism). But such reactions were attributed only to those who erroneously believed that it was real news programming (Balmas 2014), which at the time was approached as something akin to thinking that the violence in TV wresting was actually real rather than staged (some people think this, but they represent a minority who are just not "in on the joke"). Other studies extolled the social benefits of fake news of this variety, concluding that a faux-journalistic style enabled writers and hosts to better question, critique, and dispel manipulative narratives being produced by mainstream journalism and political campaigns (Jones 2010).

And it was not just "fake news" that was viewed—in some cases—as being socially beneficial. Some older work in the field of psychology more broadly found that false beliefs can have a functional value in promoting individual well-being, and thus, that false information is sometimes "more effective than correct information" in producing positive social attitudes (Nunnally and Bobren 1959). Today, misinformation is almost universally approached as a problem that needs to be solved—or at a minimum, contained. As Torabi Asr and Taboada (2019, 1) write, "Fake news is a problem. It is a big data problem." While there are good reasons for this more negative view of false information, we should appreciate it as a byproduct of both our changing social environment as well as conceptual evolution in the literature.

Conceptual Categories and Distinctions

The misinformation literature is replete with a wide variety of seemingly synonymous—but in many cases, conceptually discrete—terms. Examples include "misinformation," "disinformation," "propaganda," "fake news," "clickbait," and "false narratives." The present section discusses scholarly efforts to make sense of this muddled conceptual vocabulary.

To begin with, the root concept of information is somewhat hard to pin down. Definitions offered in the late 20th century held a somewhat narrow view that something can only count as information if it is actually true (Dretske 1981). This effectively meant that mis- and dis-information were not themselves understood as a form of information, but something else entirely. Explaining this point with a bit of analytical flare, Dretske (1983, 57) asserts that "false information, misinformation, and (grimace!) disinformation are not varieties of information—any more than a decoy duck is a kind of duck."

More recently, however, the truth requirement has grown less common as a defining attribute of "information." Fallis (2015), for instance, defines information less restrictively as "something that *represents* some part of the world as being a certain way." This revised definition allows non-true representations (e.g., mis/disinformation) to now be treated as sub-types of information. While computer scientists tend to abjure getting bogged down in these definitional debates, the treatment of misinformation as a sub-type of information has largely become standard practice in the field. Vosoughi et al (2018, 1146) make this explicit, conceptualizing misinformation as a form of "information that is inaccurate or misleading."

Misinformation, then, is now most commonly understood simply as being "false or inaccurate information" (Karlova and Fisher 2013; Floridi 2015; Shao et al 2016). Other definitions of the term, however, read in an added feature of non-intentionality—defining misinformation as "wrong information [that is published] without meaning to be wrong [and without] having a political purpose in communicating false information" (Benkler et al 2018, 24). This overlaps with an often drawn distinction between *misinformation* and *disinformation*, which will be discussed below.

Even among scholars who share a common definition for "misinformation," there is sometimes disagreement regarding the appropriate level of analysis to study and operationalize the concept. Vosoughi et al (2018) identify the misinformation unit-of-analysis as the individual news story or news-relevant post. Zhou and Zhang (2007) argue for an even more granular conception because stories may combine factual and non-factual content—so they suggest looking at smaller units-of-analysis like paragraphs. Grinberg et al (2019) go the other direction, identifying misinformation at the level of publishers according to the ratio at which they produced verifiable versus fabricated content. Lazer et al (2018) and Acerbi (2019) similarly look to the publisher-level (websites) to operationalize instances of misinformation and does not differentiate between individual articles published by each website. Ruths (2019) argues that attentiveness to levels of analysis is critical, as it can help us understand and reconcil ostensibly inconsistent findings in empirical studies of online misinformation. To this end, Ruths (2019, 348) conceives of the process of misinformation as having five key elements, each of which merit empirical scrutiny: publishers, authors, articles, audience, and rumors.

At each level of analysis, there are a variety of misinformation sub-types (or related, but non-nested, concepts) that may warrant being considered separately. Barclay (2018, 1), for instance, notes that “fake news, lies, rumors, fibs, [and] propaganda are all synonyms” but also makes a few of conceptual distinctions: *propaganda* is intentional and serves a specific agenda; *fake news* is a higher order concept that “includes propaganda, but may also include [other forms of] misinformation.” Born and Edgington (2017) make a somewhat similar distinction between *misinformation* (untrue, but not intentionally), *disinformation* (intentionally untrue), and *propaganda* (potentially true, but intentionally manipulative).

These distinctions are in some cases subtly made, and in other cases they get implicitly erased as different concepts get folded into one another. The remainder of this section will review a series of misinformation concepts (or subtypes) separately—but to be clear, it does not take a position on whether distinguishing or blending them is the better route.

While misinformation broadly encompasses “inaccurate information” (perhaps unintentionally so), disinformation that diffuses through social networks is both inaccurate *and deliberately deceptive* (Karlova and Fisher 2013; Kumar and Geethakumari 2014; Søe 2017). Disinformation is often produced textually, but it also includes visual content as well like memes, doctored videos/photos (Fallis 2009). Some studies treat disinformation explicitly as a subtype of misinformation, that which is produced intentionally “to mislead people with an agenda” in mind (Ha et al 2021). Fallis (2015) includes a pertinent caveat, however: while a piece of disinformation must be disseminated with the intent to mislead, it does not have to actually mislead its consumers— success is not a requirement to apply the concept. Even so, disinformation can be uniquely harmful because of its deliberateness—eroding trust in those exposed to it and “thereby inhibiting our ability to effectively share information with one another” (Fallis 2015, 402).

Propaganda is often treated distinctly from disinformation because it does not necessarily need to be untrue. Guess and Lyons (2020, 11), for instance, distinguish between misinformation and propaganda on the basis that the latter can be true but is always intended to disparage opposing viewpoints and “persuade people to support one political group over another.” In this sense, propaganda is not defined by the degree of truth/falsehood in a given piece of information, but instead by its purposeful usage to influence public options/actions with a political agenda in mind (Barron-Cedeno et al 2019). Information and disinformation alike could be used in propaganda.

Like propaganda, rumors have an ambiguous relationship to truth in that they may be ultimately proven accurate, inaccurate, or something in between (Friggeri et al 2014). But they are distinct from propaganda in two important ways. First, rumors are not necessarily spread deliberately or maliciously. Second, they occupy a temporary limbo between information and misinformation in that—so long as they remain rumors—they are as yet unverified (Zubiaga et al 2016). Conceptually, propaganda continues to be propaganda even if it verified as true or untrue. The rumor, once debunked or confirmed, it is better understood as misinformation, disinformation, information, or propaganda.

Fake news is often conceptualized as a sub-type of misinformation, one made distinctive by is mimicry of the journalistic process. Ha et al (2021, 291) make this nesting of concepts overt, writing “fake news is one type of misinformation presented as news.” Rochlin (2017, 388) makes a small shift toward intentionality that equates “fake news” more as a subtype of *disinformation*, asserting that fake news is a “*knowingly false* headline and story written... to look like” real news.

Lazer et al (2018, 1094) are somewhat distinctive in identifying fake news by the process through which it was produced, noting that:

> “Fake news mimics news media content in form but not in organizational process or intent. Fake-news outlets... lack the news media’s editorial norms and processes for ensuring the accuracy and credibility of information.”

This definition leads Lazer et al (2018) to associate fake news with both *misinformation* and *disinformation*, since the lack of editorial norms may generate content that is false either unintentionally or intentionally. Grinberg et al (2019, 374) adopt this same definition, and in doing so provide the helpful clarification that when fake news is conceptualized in a process-based way, “the attribution of ‘fakeness’ is... not at the level of the story but at the level of the publisher.”

Taking the literature’s proclivity for sub-typing one step farther, some work differentiates between different categories of fake news—according to whether it is created for profit, to serve a political agenda, or for social commentary (Barclay 2018). Others have gone in the opposite direction, decrying that persistent debates over what constitutes “fake news” are unnecessary distractions that bog down empirical research (Weeks and Gil de Zu´niga 2021). Still others contend˜ that the term “fake news” has lost its analytical value because it is now “irredeemably polarized” and conceptually stretched to denote any information that is not politically convenient to whoever deploys the term (Vosoughi et al 2018).

Clickbait is a distinctly online form of low-quality information, one very closely related to “fake news.” While clickbait (content whose main purpose is to attract attention) is often implicated in the spread of rumors and misinformation, this content is not necessarily false or untrue itself (Chen et al 2015). In this way, clickbait is conceptually closer to definitions of “fake news” that emphasize poor editorial norms rather than false content *per se*. The proliferation of clickbait online has been linked to shifting incentives that favor raw page views over journalistic accuracy (Chen et al 2015; Rochlin 2017). And because media business models reward minimizing the costs of producing a story but maximizing online engagement with it, publishing unverified (or completely fabricated) information has little downside. As a result, spreading misinformation and disinformation is a logical byproduct of publishers' value for clickbait journalism (Ball 2017).

Misleading, untrue, and deceptive information online need not be based wholly based on lies. In many cases, reinterpretations of conventional or scholarly information that is itself verifiable can be used produce wildly inaccurate stories. This usage of factual information to spin and peddle (mis/dis)information is conceptualized as constructing a false narrative (Introne et al 2018). The expansion of false narratives, sometimes referred to as "pseudo-knowledge" (Introne et al 2017; Froehlich 2019) suggests that the relationship between misinformation and factual information is likely more complex than is often presumed (Introne et al 2018, 1).

This intuition has been incorporated into some facets of the literature, which notably highlight that misinformation need not be wholly—or perhaps even primarily—untrue. Ha et al (2021), for instance, point out that "fake news" content regularly blends real and fake information, which problematizes the use of a hard-and-fast fake versus real dichotomy in misinformation research. This sort of fake/real blending is precisely what motivates Zhou and Zhang (2007) to operationalize misinformation at the level of individual paragraphs in an article rather than treating articles themselves as a single unit-of-analysis. In terms of conceptual categories, one option raised to deal with this complexity is that of true disinformation—or information that is deliberately misleading through the process of emphasizing only certain facts (Fallis 2015). For computer scientists, the potential existence of "true disinformation" suggests a need to detect mis/disinformation not just on the basis of textual content (which may indeed be true), but also through linguistic implicatures which may package or spin content in a way that is patently misleading (Søe 2017, 322).

1.2 **The Post-Truth Era**

Misinformation in its various forms—fake news, rumor, conspiracy, propaganda, or just plain bullshit—are widely viewed as playing an increasingly prominent view in online space and in American society writ large. This has led some observers to herald the onset of a new "post-truth" era (Benkler et al 2018). This section reviews literature on the nature and extent of the post-truth problem and discusses how this problem has been incorporated into misinformation studies.

What is Post-Truth?

To say that we are in a "post-truth era" implies that there is some macro-level attribute of "truthiness" or "non-truthiness" that can provide a meaningful—if highly general—description for a society. Some of the literature discussing post-truth in the modern age seems to genuinely have this implication in mind. Michel Foucault's (1977) work lends itself well as an intellectual origin for macro-level descriptions of truth, non-truth, and post-truth eras. Foucault's concept of *regimes of truth*, for instance is broadly defined as:

> "The types of discourse [a society] harbors... the mechanisms and instances which enable one to distinguish true from false statements... the techniques and procedures which are valorised for obtaining truth" (1977, 13).

A post-truth era, in this sense, may be understood as a involving a shift in societal discourse and methods for distinguishing truth from non-truth which either: (1) makes the discovery of truth less relevant; or (2) facilitates the heightened spread of untrue information. Post-truth is often defined in a way that corresponds to the presumption that such a shift has indeed occurred.

In 2016, "post-truth" was selected as Oxford Dictionaries' word of the year, and it was defined as "circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief." On this definition, the discovery of truth is less relevant because objective facts have diminished in influence and also because emotions have become more influential in public efforts to discriminate between true and false information. Visvisi and Lytras (2019, 2) frame their discussion of technology in the post-truth era by defining this era as one in which "appearances are given priority over objectivity... and emotional and subjective assessments... cloud the essence of things and so the truth." This emphasis on emotional evaluations of information is broadly consistent among those who raise concern about an age of post-truth. Ha et al (2021, 293), for instance, treat post-truth as a highly emotion-based information market.

The value that emotional appeals have over objective or verifiable information has been used to explain Donald Trump's success in the 2016 election as well as the success of the Brexit referendum in the U.K.(Oyserman and Dawson 2021). And research on computer science attentive to post-truth has been growing increasingly worried that the privilege afforded to emotional and identity-based reasoning renders individuals more vulnerable to misinformation and cognitive manipulation (Ismailova et al 2020).

Expressing a concern that a widespread devaluing of truth influences public behavior, Strong (2016-2017, 138) uses the term "post-truth society" to describe a system in which "individuals and institutions adopt behaviors or beliefs that are patently at odds with observable reality." Gorrell et al (2019) express a similar concern, with a key difference being that they use the term "posttruth" as an adjective describing a form of politics (i.e., "post-truth politics") in which parties, governments, and politicians frame key public policy issues around propaganda instead of facts.

There are, of course, a few more charitable—and less cynical—interpretations of individuals who occupy a space of "post-truth" reality. Fuller, for example, describes the "post-truther" not as someone who outright denies the existence of facts or deliberately peddles lies—but instead as someone who simply wishes to "dispel the mystery in which the creation and maintenance of facts tend to be shrouded" (2018, 19). As a philosopher, Fuller's reading of post-truth is not one of an age in which truth itself has lost its value, but instead one in which the dominant regime of truth (Foucault 1977) is currently being contested.

What Does the Post-Truth Era Look Like?

The post-truth era as it looks in practice is often characterized as taking on a number of distinctive features associated with heightened skepticism about the veracity of information circulated on- and offline. First, open data—particularly when provided by governments and agencies with partisan leadership—is increasingly unreliable (Colborne and Smit 2017). Second, individuals are growing more distrustful of political and media institutions as well as more uncertain about the accuracy of the information they are provided (Flintham et al 2018; Wagner and Boczkowski 2019). Third, the truth value of information no longer matters (or it matters much less), as many individuals now treat "fake news" as any information that challenges their pre-existing beliefs of preferences (Rochlin 2017). Fourth, and relatedly, is that virtually no information can be treated as objective or impartial by its consumers (Waisbord 2018). Fifth, because emotion is the currency of the day, the use of moralistic or emotional framing is key for facilitating the spread of information in online social networks (Brady et al 2017)

Post-Truth Era as Misinformation Era

In scholarship on computer science, the post-truth era is often equated with being an age of rampant misinformation (Mocanu et al 2015). This relationship was very notably magnified by the World Economic Forum's 2013 decision to list "massive digital misinformation" as one of the most critical threats to modern society. Recent work on algorithmic methods for detecting online misinformation continues to equate the rapid increase in misinformation with the presumed dawn of a post-truth era (Popat et al 2018, 155). Still others warn that the post-truth era presents acute misinformation vulnerabilities that should not escape the attention of scholars and policymakers (Ismailova et al 2020), particularly by jeopardizing the quality of data available for big data analysis in ways that are difficult to detect and mitigate (Colborne and Smit 2017, 2588).

To meet the challenges of the post-truth era (one increasingly understood as a misinformation era), computer scientists have been hard a work designing a variety of methods to automate the detection of different forms of online misinformation: *Truthy* for non- or partially-truthful information (Ratkiewicz et al 2011); *Hoaxy* for misinformation broadly (Shao et al 2016); *Proppy* for online propaganada (Barron-Cedeno et al 2019); the *Rumor Tracking System* (RTS) for rumor verification (Samreen et al 2020); the *PHEME-Project* also for rumor verification (Zubiaga et al 2015); and *CredEye* for generating information credibility assessments (Popat et al 2018).

The Origins of Post-Truth

If we are indeed in a post-truth era, what got us here in the first place? To begin with a crude timeline, the post-truth age itself is sometimes viewed as reaching full force in 2016—though some would trace its roots back farther. Matthew D'Ancona, an English journalist, begins his book *Post-Truth* by declaring:

> "To every thing there is a season: 1968 marked the revolution in personal freedom and the yearning for social progress; 1989 will be remembered for the collapse of totalitarianism; and 2016 was the year that definitively launched the era of 'Post-Truth"' (2017, 1).

Lewandowsky et al (2017, 2) contend that the post-truth world is emerging as a result of megatrends unfolding in American society: "the decline in social capital, growing economic inequality, increased polarization, declining trust in science, and an increasingly fractionated media landscape." These factors facilitated the spread of misinformation, made individuals more vulnerable to it, and also made it more difficult to effectively contain. Gorrell et al (2019) add in the factor of technological change, arguing that the "lowered bar to publication offered by Web 2.0" facilitated the rise of post-truth in virtual space. Acerbi (2019) postulates that misinformation has a unique comparative advantage over factual information when competing for social media users attention. This is because misinformation is not constrained by reality, and so it can be easily shaped to be more cognitively appealing, attention grabbing, sensational, and sectionrable.

Benkler et al (2018) list a standard array of usual suspects thought to be accelerating the turn to post-truth in American society: fake news entrepreneurs, Russian hackers and bots, social media companies, alt-right trolls, right wing media, mainstream media, and Donald Trump himself. Ultimately, they conclude that the prime accelerant is not any one of these actors, but instead a different mega-trend of rising political polarization and partisan insulation (Benkler et al 2018, 23). For their part, journalists appear disproportionately likely to place the blame for post-truth primarily upon the media itself and its evolving business model—one which rewards lower investments in high-quality reporting and allows publishing of false (or unchecked) content to be highly profitable (Ball 2017).

Each of these factors may have played a contributing role to the rapid proliferation of misinformation and the progressive devaluing of objective facts. Even so, some research on the post-truth era cautions against being too apocalyptic in our assessments, and further, to think twice before concluding that a Pandora's box of post-truth has been irrevocably opened. Lewandowsky et al (2017, 3), for instance, find that while American society moving in the direction of a post-truth age, this distopian future is still just "a possible future." Lee McIntyre's philosophical account of post-truth views this phenomenon as being less about reality than it is about "the way that humans *react* to reality," which suggests that post-truth is a state that people can choose—or refuse—to occupy (2018, 172). Finally, Corner (2017, 1106) finds that post-truth alarmism can be useful in some ways but is also by-and-large quite reductionist and prone to understate the extent to which "truth" still matters for a wide array of public policy issues.

References

[1] Acerbi, Alberto. (2019). “Cognitive Attraction and Online Misinformation.” *Palgrave Communications* 5(1): 1–7.

[2] Allport, Gordon and Leo Postman. (1947). *The Psychology of Rumor*. New York: Holt, Rinehart, and Winston, Inc.

[3] Aristotle. (2007). Translation by George A.Kennedy. *On Rhetoric: A Theory of Civic Discourse*. New York: Oxford University press.

[4] Ball, James. (2017). *Post-Truth: How Bullshit Conquered the World*. London: Biteback Publishing.

[5] Baym, Geoffrey. (2005). “The Daily Show: Discursive Integration and the Reinvention of Political Journalism.” *Political Communication* 22(3): 259–276.

[6] Balmas, Meital. (2014). “When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism.” *Communication Research* 41(3): 430–454.

[7] Barron-Cedeno, Alberto, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. (2019). “Proppy: A System to Unmask Propaganda in Online News.” *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1): 9847–9848.

[8] Benkler, Yochai, Robert Farris, and Hal Roberts. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. New York: Oxford University press.

[9] Born, Kelly and Nell Edgington. (2017). “Analysis of Philanthropic Opportunities to Mitigate the Disinformation/Misinformation Problem.” *Hewlett Foundation*, https://www.hewlett.org/wp-content/uploads/2017/11/ Hewlett-Disinformation-Propaganda-Report.pdf.

[10] Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. (2017). “Emotion Shapes the Diffusion of Moralized Content in Social Networks.” *Proceedings of the National Academy of Sciences* 114: 7313–7318.

[11] Chen, Xinran, and Sei-Ching Joanna Sin. (2013). “‘Misinformation? What of It?’ Motivations and Individual Differences in Misinformation Sharing on Social Media.” *Proceedings of the American Society for Information Science and Technology* 50(1): 1–4.

[12] Chen, Yimin, Niall J. Conroy, and Victoria L. Rubin. (2013). "Misleading Online Content: Recognizing Clickbait as 'False News'." *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*: 15–19.

[13] Colborne, Adrienne, and Michael Smit. (2017). "Identifying and Mitigating Risks to the Quality of Open Data in the Post-Truth Era." *IEEE International Conference on Big Data (Big Data)*: 2588–2594.

[14] Corner, John. (2017). "Fake News, Post-Truth, and Media-Political Change." *Media, Culture, & Society* 39(7): 1100–1107.

[15] D'Ancona, Matthew. (2017). *Post-Truth: The New War on Truth and How to Fight Back.* New York: Random House.

[16] Day, Amber and Ethan Thompson. (2012). "Live from New York, It's the Fake News! Saturday Night Live and the (Non)politics of Parody." *Popular Communication* 10(1-2): 170–182. [17] Dretske, Fred I. (1981). *Knowledge and the Flow of Information.* Cambridge: MIT Press.

[18] Dretske, Fred I. (1983). "Precis of Knowledge and the Flow of Information."´ *Behavioral and Brain Sciences* 6(1): 55–90.

[19] Fallis, Don. (2009). "A Conceptual Analysis of Disinformation" In *iConference Proceedings*: https://www.ideals.illinois.edu/bitstream/handle/2142/15205/fallis_ disinfo1.pdf?seq.

[20] Fallis, Don. (2015). "What Is Disinformation?" *Library Trends* 63(3): 401–426.

[21] Floridi, Luciano. (2015). "Semantic Conceptions of Information." In Stanford Encyclopedia of Philosophy, https://seop.illc.uva.nl/entries/information-semantic/.

[22] Foucault, Michel. (1977). "The Political Function of the Intellectual." *Radical Philosophy* 7: 12–14.

[23] Flintham, Martin, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. (2018). "Falling for Fake News: Investigating the Consumption of News via Social Media." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*: 1–10.

[24] Friggeri, Adrien, Lada Adamic, Dean Eckles, and Justin Cheng. (2015). "Rumor Cascades." *Proceedings of the International AAAI Conference on Web and Social Media* 8(1).

[25] Froehlich, Thomas J. (2019). "The Role of Pseudo-Cognitive Authorities and Self-Deception in the Dissemination of Fake News." *Open Information Science* 3: 115–136.

[26] Fuller, Steve. (2018). “What Can Philosophy Teach Us About the Post-Truth Condition?” In *Post-Truth, Fake News: Viral Modernity & Higher Education* , edited by Michael A. Peters, Sharon Rider, Mats Hyvonen, and Tina Besley. New York: Springer.¨

[27] Gorrell, Genevieve, Mehmet E. Bakir, Ian Roberts, Mark A. Greenwood, Benedetta Iavarone, and Kalina Bontcheva. (2019). “Partisanship, Propaganda, and Post-Truth Politics: Quantifying Impact in Online Debates.” *ArXiv preprint*.

[28] Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. (2019). “Fake News on Twitter during the 2016 US Presidential Election.” *Science* 363(6425): 374–378.

[29] Guess, Andrew M.and Benjamin A.Lyons. (2020). “Misinformation, Disinformation, and Online Propaganda.” In *Social Media and Democracy: The State of the Field and Prospects for Reform*, edited by Nathaniel Persily and Joshua A.Tucker. New York: Cambridge University Press.

[30] Ha, Louisa, Loarre Andreu Perez, and Rik Ray. (2021). “Mapping Recent Development in Scholarship on Fake News and Misinformation, 2008 to 2017: Disciplinary Contribution, Topics, and Impact.” *American Behavioral Scientist* 65(2): 290–315.

[31] Holbert, R.Lance. (2005). “A Typology for the Study of Entertainment Television and Politics.” *American Behavioral Scientist* 49(3): 436–453.

[32] Introne, Joshua, Luca Iandoli, Julia DeCook, Irem Gokce Yildirim, and Shaima Elzeini. (2017). “The Collaborative Construction and Evolution of Pseudo-Knowledge in Online Conversations.” *Proceedings of the 8th International Conference on Social Media & Society*: 1–10.

[33] Introne, Joshua, Irem Gokce Yildirim, Luca Iandoli, Julia DeCook, and Shaima Elzeini. (2018). “How People Weave Online Information into Pseudoknowledge.” *Social Media + Society* 4(3): 1–15.

[34] Ismailova, Larisa, Viacheslav Wolfengagen, Sergey Kosikov, Mikhail Maslov, and Juliane Dohrn. (2020). “Semantic Models to Indicate Post-Truth with Fake News Channels.” *Procedia Computer Science* 169: 297–303.

[35] Jones, Jeffrey P. (2010). *Entertaining Politics: Satiric Television and Political Engagement*. Lanham: Rowman & Littlefield.

[36] Karlova, Natascha and Karen Fisher. (2013). “A Social Diffusion Model of Misinformation and Disinformation for Understanding Human Information Behavior.” *Information Research* 18(1).

[37] Knapp, Robert H. (1944). “The Psychology of Rumor.” *Public Opinion Quarterly* 8(1): 22– 37.

[38] Kumar, KP Krishna and G.Geethakumari. (2013). “Detecting Misinformation in Online Social Networks Using Cognitive Psychology.” *Human-centric Computing and Information Sciences* 4(1): 1–22.

[39] David, L., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D. and Schudson, M. (2018). “The Science of Fake News: Addressing Fake News Requires a Multidisciplinary Effort.” *Science* 359(8).

[40] Lewandowsky, Stephan, Ullrich KH Ecker, and John Cook. (2017). “Beyond Misinformation: Understanding and Coping with the ‘Post-Truth’ Era.” *Journal of Applied Research in Sectionry and Cognition* 6(4): 353–369.

[41] McIntyre, Lee. (2018). Translation by Robin Waterfield. *Post-Truth*. Cambridge: MIT Press.

[42] Mocanu, Delia, Luca Rossi, Qian Zhang, Marton Karsai, and Walter Quattrociocchi. (2015). “Collective Attention in the Age of (Mis) Information.” *Computers in Human Behavior* 51: 1198–1204.

[43] Nunnally, Jum C.and Howard M.Bobren. (1959). “Attitude Change with False Information.” *Public Opinion Quarterly* 23(2): 260–266.

[44] Oyserman, Daphna and Andrew Dawson. (2021). “Your Fake News, Our Facts: IdentityBased Motivation Shapes What We Believe, Share and Accept.” In *The Psychology of Fake News*, edited by Rainer Greifeneder, Mariela Jaffe, Eryn Newman, Norbert Schwarz. Milton Park: Routledge Press.

[45] Plato. (1994). Translation by Robin Waterfield. *Gorgias*. New York: Oxford University Press.

[46] Plato. (2000). Translation by Tom Griffith. Edited by G.R.F.Ferrari. *The Republic*. New York: Cambridge University press.

[47] Popat, Kashyap, Subhabrata Mukherjee, Jannik Strotgen, and Gerhard Weikum. (2018).¨ “‘CredEye: A Credibility Lens for Analyzing and Explaining Misinformation.” *Companion Proceedings of the The Web Conference*: 155–158.

[48] Qazvinian, Vahed, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. (2011). “Rumor Has It: Identifying Misinformation in Microblogs.” *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*: 1589–1599.

[49] Ratkiewicz, Jacob, Michael Conover, Mark Meiss, Bruno Gonc¸alves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. (2011). “Truthy: Mapping the Spread of

Astroturf in Microblog Streams." *Proceedings of the 20th International Conference Companion on World Wide Web*: 249–252.

[50] Rochlin, Nick. (2017). "Fake News: Belief in Post-Truth." *Library Hi Tech* 35(3): 386–392.

[51] Ruths, Derek. (2019). "The Misinformation Machine." *Science* 363(6425): 348–348.

[52] Samreen, Alia, Adnan Ahmad, and Furkh Zeshan. (2020). "Searching for Truth in the Post-Truth Age." *3rd International Conference on Advancements in Computational Sciences (ICACS)*: 1–5.

[53] Shao, Chengcheng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. (2016). "Hoaxy: A Platform for Tracking Online Misinformation." *Proceedings of the 25th international conference companion on world wide web*: 745–750.

[54] Søe, Sille Obelitz. (2017). "Algorithmic Detection of Misinformation and Disinformation: Gricean Perspectives." *Journal of Documentation* 74(2): 309–332.

[55] Strong, S.I. (2016–2017). "Alternative Facts and the Post-Truth Society: Meeting the Challenge." *University of Pennsylvania Law Review* 165: 137–146.

[56] Torabi Asr, Fatemeh and Maite Taboada. (2019). "Big Data and Quality Data for Fake News and Misinformation Detection." *Big Data & Society* 6(1): 1–14.

[57] Visvizi, Anna and Miltiadis D.Lytras (eds.). (2018). *Politics and Technology in the Post-Truth Era*. Bingley: Emerald Publishing.

[58] Vosoughi, Soroush, Deb Roy, and Sinan Aral. (2018). "The Spread of True and False News Online." *Science* 359(6380): 1146–1151.

[59] Wagner, Maria Celeste and Pablo Boczkowski. (2019). "The Reception of Fake News: The Interpretations and Practices that Shape the Consumption of Perceived Misinformation." *Digital Journalism* 7(7): 870–885.

[60] Waisbord, Silvio. (2018). "The Elective Affinity between Post-Truth Communication and Populist Politics." *Communication Research and Practice* 4(1): 17–34.

[61] Weeks, Brian E., and Homero Gil de Zu´niga. (2021). "What's Next? Six Observations for˜ the Future of Political Misinformation Research." *American Behavioral Scientist* 65(2): 277–289.

[62] Zhou, Lina, and Dongsong Zhang.. (2015). "An Ontology-Supported Misinformation Model: Toward a Digital Misinformation Library." *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 37(5): 804–813.

[63] Zubiaga, Arkaitz, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. (2015). "Towards Detecting Rumours in Social Media." *arXiv preprint*.

[64] Zubiaga, Arkaitz, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. (2016). "Analyzing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads." *PloS One* 11(3).Geraldine Wong Sak Hoi, and Peter Tolmie

2. The Diffusion of Misinformation Online

Prepared primarily by SS

This section addresses the diffusion of misinformation online in three steps. First, it discusses research on how misinformation competes with accurate information to reach consumers and shape their beliefs/attitudes/knowledge. Second, it turns to scholarship on misinformation diffusion patterns to better understand when and how false content gets disseminated throughout online social networks. Third, it concludes by scrutinizing whether certain types of social networks are more conducive to the spread of online misinformation.

Each topic will be discussed sequentially. The goal of this structure is to: (1) begin by identifying a problem—that misinformation is shaping up to be a formidable opponent for reliable information when competing to attract and influence internet users; and (2) then proceed to clarify the nature of the problem by examining how misinformation spreads online as well as which social network attributes facilitate that spread.

2.1 Information vs Misinformation in Online Space

Users of online social media are situated within an environment in which good information and misinformation compete to attract their interests and influence their beliefs, often in disparate directions (Budak et al 2011; Weng et al 2012; Wang et al 2014). Spreading online misinformation is sometimes even deployed as a deliberate tactic for distracting individuals and reducing their attentiveness to real events occurring offline in the real world, as King et al (2017) conclude in their study on why the Chinese government spreads fabricated information online. Accordingly, the capacity for misinformation to present itself as a viable competitor against verifiable information is a rising problem in the modern period.

Scholarship on the competition between information and misinformation has produced troubling results that suggest the latter may be getting the upper hand. In a study on how different types of information are consumed by a sample of 1.2 million Facebook users, Bessi et al (2015) show that conspiracy news is more likely to be shared online than scientific news. Vosoughi et al (2018) analyze the spread of 126,000 true and false news stories distributed on Twitter between 2006–2017, notably finding that false news diffuses faster, farther, and more broadly throughout online social networks than true news. Through a study of 2,691 Weibo posts in China related to gynecologic cancers between 2015–2016, Chen et al (2018) similarly find that misinformation diffused significantly more broadly and deeply than true information. Ackland and Gwynn (2021) analyze the diffusion of 14 news stories on Twitter (one of which was untruthful) on three randomly selected days during 2017—May 22, June 16, and July 4. Their results suggest that true information retains an edge over misinformation, as the untruthful story in their sample reached fewer people, spread slower, and diffused less broadly throughout the network.[1]

The optimistic conclusions provided in Ackland and Gwynn (2021) appear to be more-or-less of a minority in the misinformation diffusion literature. A body of recent work continues to conclude that misinformation: (1) spreads so rapidly online that attempts at manual verification are inherently quite limited in their effectiveness (Popat et al 2018);[2] (2) cascades deeper throughout social networks than does previously verified information (Friggeri et al 2014);[3] The and (3) reverberates online at least as long as accurate information (Mocanu et al 2015).[4]

[1] Given data limitations, however (selecting only one untrue story, for a very limited time period), Ackland and Gwynn (2021, 44) present their contribution as being primarily methodological "rather than a definitive test of the diffusion behavior of true and false news.

[2] Popat et al (2018) devise CredEye, an automated system for assessing the credibility of information online, precisely because they view manual verification as unscalable.

[3] Friggeri et al (2014) analyze the diffusion of 4,761 rumors on Facebook, distinguishing between rumors that are true and false using classifications made by Snopes.com. They conclude that compared to a non-rumor reference sample, rumors (both true and false) cascade more deeply within the Facebook network.

[4] Mocanu et al (2015) study a sample of 2.3 million Facebook users to assess how individuals consumed different types of information during the 2013 Italian electoral campaign. The authors find that posts containing unsubstantiated information and those with verifiable information display: (1) a highly similar pattern of user

The conclusions from Friggeri et al's (2014) study of rumor diffusion on Facebook indicate that information of dubious accuracy (i.e., *rumors*) tends to get spread more widely on social media than reliable information (i.e., *non-rumors*), though the authors also find that rumors which eventually turn out to be true still generally outperform those that are later proven false. Sommariva et al's (2018) analysis of Zika-related news stories circulating on Facebook, LinkedIn, Twitter, Pinterest, and GooglePlus between February 2016 to January 2017 supports the conclusion that rumors diffuse more broadly online than trustworthy information. Indeed, Sommariva et al (2018, 246) find that Zika "rumors had three times more shares than verified stories." Zubiaga et al (2016) demonstrate that false rumors spreading online take longer to verify than true rumors (likely due to the heightened difficulty of proving a negative), and further, that social media users tend to default to believing any type of rumor (true or false) so long as it remains unverified. These conclusions are derived from Zubiaga et al's (2016) Twitter analysis of the lifecycle for 330 rumor threads (n=4,842 tweets) associated with nine newsworthy events. The heightened diffusion of rumors in only social networks is partly attributed to the observation that social media users do not exert much effort to share debunks of a false rumor, while users who continue to propagate a false rumor remain more active in comparison (Friggeri et al 2014; Zubiaga et al 2016).

In their analysis of misinformation diffusion on Twitter, Shao et al (2018a) find that the total amount fact-checking tweets is absolutely dwarfed by the total amount of misinformation tweets by users (at about a 1:17 ratio).[5] Moreover, fact-checking typically lags behind misinformation by

engagements; and (2) a highly similar post lifetime, measured the temporal distance between the first and last comments of the post.

[5] Shao et al (2018a) conduct this analysis using Hoaxy, an open source system that they designed to collect "a full and comprehensive set of tweets that share (i.e., include a link to)" a misinformation article or fact check starting in 2016. The authors emphasize that "Hoaxy collects 100% of these tweets, not a sample."

about one day—giving false stories healthy first-mover advantage which allows them to grow more popular than the corresponding debunking (Shao et al 2018b).[6]

And after online misinformation is corrected, it still exhibits a lingering effect as individuals (even those who know about and recall the corrections) continue to rely on it to formulate their beliefs. Lewandowsky (2005) demonstrates the enduring effect of previously debunked misinformation through a survey of American (n=302), Australian (n=158), and German (n=412) university community members' beliefs on the Iraq War. The results of Lewandowsky's (2005, 193) surveys indicate that American respondents did not update their beliefs about misinformation items related to the Iraq War even when they had knowledge that such information had previously been corrected or retracted.

Even studies that identify fact-checking as being fairly effective in combatingonline misinformation observe that the efficacy of corrective information is limited because of: (1) the sheer amount of online misinformation on different topics; and (2) the short-term spread of fact-checks compared to the longer durability of misinformation. Burel et al (2020) demonstrate the limited— but still real—utility of fact-checking through an analysis of COVID-19 misinformation and its corresponding fact-checks on Twitter between December 2019 and May 2020. The authors find that "fact-checked information is shared less often than the corresponding misinformation URLs" (Burel et al 2020, 8) and, further, that while fact-checking does reduce the spread of COVID-19 misinformation in the short-term, "the reduction... is mostly temporary, [which] indicates that the reduction power of fact-checking is impeded by its inability to be shared over long period of time" (Burel et al 2020, 12).

Depicting a fairly balanced competition between information and misinformation in the experimental setting (MTurk), van der Linden et al (2017) show that a persuasive fact's effects on individual beliefs gets negated when it is accompanied a piece of misinformation—suggesting that misinformation can cancel out fact-based information. This cancellation effect has been

[6] Shao et al (2018b) identify this one-day lag between misinformation and fact-checking by analyzing the diffusion of tweets recorded through the Hoaxy system over time.

observed to be a deliberate tactic used by misinformation purveyors outside of the experimental setting. Shahi et al (2021), for instance, note that misinformation tweets about COVID-19 were commonly distinctive in their heightened concern with discrediting other sources of information circulating on social media.[7]

A Misinformation Advantage? Sources and Descriptive Statistics

As truthful and untruthful information vie against one another in online space, accurate information can be “easily crowded out by sensationalized news” that is either unverified or even demonstrably false (Wang et al 2019, 1). Indeed, some scholars offer good reasons for why misinformation may actually have an advantage as it competes with accurate information to influence social media users. Psychological sources of this advantage include the average individual’s: (1) default presumption that unverified information is true until proven otherwise (Zubiaga et al 2016; Schwarz and Jalbert 2021);[8] (2) inability to differentiate between information and misinformation due to limited cognitive capacities interacting with information overload online (Qiu et al 2017);[9] (3) value for belief *consistency* over belief *accuracy* (Kuklinski et al

[7] Shahi et al (2021) analyze COVID-19 misinformation on Twitter by collecting “all tweets mentioned in the verdicts of fact-checked claims related to COVID-19 by over 92 professional fact-checking organisations between January and mid-July 2020.” By analyzing the content of these Tweets, Shahi et al (2021) show that the authors of misinformation use “less tentative language” and are “more often concerned with discrediting other information on social media.”

[8] In reviewing findings from psychology work on how individuals evaluate information, Schwarz and Jalbert 2021 observe that “*Going beyond the default of information acceptance requires motivation and cognitive resources, which we are most likely to invest when the topic is important to us and there are few competing demands and distractions. In the absence of these conditions, information is likely to be accepted—and sometimes passed on—without much scrutiny.*”

[9] Qiu et al (2017) devise a simple theoretical model of an online social network in which individual users: (1) prefer high quality information; but (2) are limited by virtue of having to manage a heavy flow of information. The authors conclude that the viral spread of low-quality information is facilitated by the limited attention of individual social network users as well as a higher level of information overload in the online market.

2000);[10] (4) higher likelihood of sharing “novel” information even if it is unreliable or false (Vosoughi et al 2018).[11]

Economic sources of a misinformation advantage have also been postulated, namely that types of misinformation like *fake news* can be both cheap to produce and highly profitable. Allcott and Gentzkow (2017, 222) collect data on the share of traffic from different sources (direct browsing, search engines, social media) for the top 690 US news websites and 65 fake websites, showing that real news outlets receive only 10.1% of interactions through social media whereas fake news sites get 41.8% of their interactions from social media. Explaining this discrepancy, Allcott and Gentzkow (2017, 221) note that social media is uniquely hospitable to creating fake news because:

> “The fixed costs of entering the market... are vanishingly small. This increases the relative profitability of the small-scale, short-term strategies often adopted by fake news producers, and reduces the relative importance of building long-term reputations for quality.”

As a journalist, James Ball (2017) also traces the expansion of low-quality news information to profit incentives for publishers. In some cases, fake or unsubstantiated news gets published “as a side-effect of shrinking newsrooms and cost cutting,” but in other cases it is based in a more cynical logic of pushing out cheap stories that can generate mass traffic whatever the consequences (Ball 2017, 142).

Search engine and social media algorithms sometimes exacerbate the misinformation problem by amplifying online misinformation and disseminating it to specific users as

[10] Kuklinski et al (2000) conduct a telephone survey of 1,160 Illinois residents, showing that initially misinformed respondents did not update their policy preferences on welfare spending even when informed with correct facts. The authors did, however, find that misinformed citizens will update their policy preferences to align with accurate facts only if those facts are “presented in a way that hits them between the eyes” by explicitly correcting stated misperceptions with clearly authoritative information (p. 805).

[11] Vosoughi et al (2018) postulate that social network users preference for sharing novel information is the reason why false news diffused faster, farther, and deeper than true news in their analysis of rumor cascades on Twitter (n=126,000 rumors) between 2006 and 2017.

recommended content. Borges et al (2019, 610) argue that Facebook's algorithms for personalizing and curating users' content recommendations "potentially creates filter bubbles and echo chambers that can lead to... perpetuation of misinformation." Hussein et al (2020) audit the degree to which YouTube's personalization algorithms (based on watch history, age, gender, geolocation) contribute to the amplification of misinformation. Their analysis concludes that filter-bubbles are created through a users previous watch history, and that "watching videos that promote misinformation leads to more misinformative video recommendations (Hussein et al 2020, 48).

If misinformation has an advantage over accurate information in virtual space, we may expect this advantage to manifest in a high frequency of user interaction with false online content. As noted in Section #1, the World Economic Forum (2013) indicated that "massive digital misinformation" is a crisis in modern society, and references to a presumed "post-truth" era have become a staple in journalism, communications, and computer science research on misinformation (Ball 2017; Colborne and Smit 2017; Waisbord 2018; Ismailova et al 2020; Ha et al 2021).

One assumption very commonly associated with these references to "post-truth" is that in online environments, false information reaches large audiences "by spreading rapidly from one individual to another" (Roozenbeek and van der Linden 2019, 570). This reflects a much broader conventional wisdom that fake news and misinformation are now becoming extremely prevalent online for a variety of topics ranging from climate change (Treen et al 2020),[12] medicine (SuarezLledo and Alvarez-Galvez 2021),[13] and politics (Gorrell et al 2019).[14]

[12] Treen et al (2020) review the literature on climate change misinformation online. They identify two key research gaps: (1) "understanding the diffusion of climate change information on social media"; and (2) "examining whether misinformation extends to climate alarmism as well as climate denial" (p.1).

[13] Suarez-Lledo and Alvarez-Galvez (2021) review 69 studies of health misinformation on social media. They find that the most prominent topics of these studies were: vaccines; drugs and smoking; noncommunicable diseases; pandemics; eating disorders; and medical treatments.

[14] Gorrell et al (2019) analyze 17.5 million tweets posted on Twitter in the lead-up to Brexit, a sample selected using the main Brexit related hashtags. Their analysis finds that non-factual and misleading claims made by

Empirical studies on the prevalence of misinformation content online—as well as internet users' exposure to it—produces mixed results when evaluating this conventional wisdom. In some studies, fears of widespread online misinformation appear well-founded. Allcott and Gentzkow (2017, 212) estimate the number of fake news stories clicked and read online during the 2016 US election period to be approximately 760 million (or three per American adult).[15] In a later study, Allcott et al (2019, 2) show that while engagement with fake news online has declined since 2016, it still remains quite high—with fake news sites receiving about two-thirds as many Facebook engagements as conventional news sites (approximately 60million fake news engagements per month).[16] Analyzing rumor cascades on Twitter between 2006–2017, Vosoughi et al (2018) find that around 126,000 different rumors were spread by nearly 3 million people. In that same analysis, Vosoughi et al (2018) further conclude that false news reaches even more people online than the truth as misinformation gets diffused more broadly.

These results are worrying and suggest a rather high prevalence of interaction with misinformation on social media. At the same time, a number of articles published over the past three years suggest that the concerns about the prevalence of online misinformation may be somewhat over-exaggerated. Guess et al (2019, 1) conclude that sharing misinformation content is "a relatively rare activity" as over 90% of individuals in their study of Facebook sharing activity did not share any stories from fake news sites.[17] Guess et al (2018) further show that during the 2016 US election period: (1) political fake news accounted for only 2.6% of American's

politicians during the referendum campaign provided the basis for a massive amount of online misinformation diffusing in the lead-up to Brexit.

[15] Allcott and Gentzkow (2017) arrive at this estimate using a database of fake news articles circulated three-months prior to the 2016 election (fake news articles were identified through Snopes.com, PolitiFact.com, and Buzzfeed). This provides them with a collection of 156 fake news articles, which they identify as having been shared 37.6 million times on Facebook. To approximate the number times each story was read (n=760 million), the authors assume that each share on Facebook is equivalent to 20 page visits.

[16] Allcott et al (2019) identify this overtime pattern by tracking the diffusion of content from 569 fake news websites as well as 9540 fake news stories on Facebook and Twitter between July 2015 and July 2018. User interactions with such fake news content was collected through BuzzSumo, a commercial content database which tracks the volume of user interactions with content on social media platforms, including Facebook and Twitter.

[17] This study by Guess et al (2019) linked an online survey of (n=3500) respondents with behavioral data on the respondents' Facebook sharing history during the 2016 US presidential campaign.

engagement with online news content; and (2) approximately 27.4% of American adults visited an article on a political fake news site.[18] With respect to Twitter, Grinberg et al (2019) find that engagement with fake news is heavily concentrated among a small segment of the online population, with only 1% of individuals accounting for over 80% of fake news interactions (and only 0.1% of users accounting for nearly 80% of fake news shares).[19] In a study examining Americans' consumption of online and television news jointly, Allen et al (2020, 1) conclude that "fake news comprises only 0.15% of Americans' daily media diet."[20] Taken together, these studies suggest that while the total number of interactions with fake news on social media may be high, the number of people actually engaging in such interactions with misinformation content remains quite low.

Outside of the US, Fletcher et al (2018) investigate the reach of online misinformation in France and Italy—concluding that only a small proportion of the population interacts with fake news websites. Their results are derived from Facebook interactions with a sample of around 300 websites in each country that independent fact-checkers identify as publishers of fake news. Most of the fake news sites in their study were accessed by less than 1% of the online population in each country, whereas the most popular conventional news websites reached approximately 22% (France) and 50% (Italy) of internet users each month.

While the bulk of research on misinformation specifically focuses on its prevalence and reach

[18] These estimates from Guess et al (2018) are based on a data set that combines "responses to an online public opinion survey from a national sample of 2,525 Americans with web traffic data collected passively from their computers with their consent during the October 7–November 14, 2016 period."

[19] These estimates from Grinberg et al (2019) are based on database linking a nationally representative sample of US voter registration records to Twitter accounts (n=16,442 accounts). The authors collected tweets sent by these accounts between 8/1/2016–12/6/2016, specifically tracking tweets that engaged with content from fake news sources (publisher level definition).

[20] Allen et al (2020) construct a nationally representative sample of television, mobile, and desktop consumption. TV consumption data comes from Nielsen's nationally representative TV panel (n=100,000). Mobile and desktop consumption come from Comscore's nationally representative sample (for mobile and desktop) as well as Nielsen's nationally representative desktop only web panel. Fake news consumption is operationalized only for online sources, and it reflects the time spent on any one of 98 websites identified by fact-checkers, researchers, or journalists as being sources of deceptive or low-quality content. The total "*media diet*" reflects the time spent consuming media of any type from any source (TV/mobile/computer).

in the United States, cross-national studies will likely continue to prove rewarding in elucidating the scale of the online misinformation problem in comparative perspective. There are, for instance, good reasons to expect that exposure to misinformation—and to specific types of misinformation— varies across countries. The Chinese government, for instance, is particularly active in using online disinformation to influence public opinion in the mainland (King et al 2017) as well as among audiences in Hong Kong and Taiwan (Myers et al 2019; Dickey 2019). And in the Arab world, the COVID-19 period coincided with an explosion of misinformation of a distinctly religious variety (false *hadiths* and *fatwas*) to the extent that Alimardani and Elswah (2020, 1) conclude that "religious misinformation is a defining characteristic of the MENA's online sphere." Accordingly, the prevalence of misinformation online may be wildly different cross-nationally, and some types of online misinformation that are widespread outside of the United States (e.g., falsified government statistics, fabricated *hadiths*) may remain undetected in studies that conceptualize engagement with misinformation with only American audiences in mind.

2.2 Models of (Mis)information Diffusion

If information and misinformation compete for user attention online, how might we model the nature of this competition? This section begins first by reviewing research that proposes theoretical models of (mis)information diffusion online. It then turns to scholarship that investigates the process of (mis)information diffusion empirically through the use of social network data.

The most commonly used models of how information spreads online can be broadly grouped into two camps: (1) epidemiological models; and (2) topological models. Each camp begins with a mathematical model that reflects the presumed (mis)information diffusion process and then identifies the parameters either through simulations on synthetic network structures, or alternatively, using real data social network structures with the goal of optimizing model fit. Before proceeding, it is important to note that

Epidemiological Models

Epidemiological models treat the spread of misinformation online as diffusing like an infectious disease. These models divide the online population into several compartments—some combination of: susceptible (S), exposed (E), infected (I), recovered (R). The goal of epidemiological models is typically to estimate the number of infected (I) individuals (i.e., those influenced by misinformation) within a fixed population over time (Garvic and Bagdasaryan 2019). Individuals can shift from one compartment to another—most notably, to-and-from the "infected" (I) compartment— according to probability parameters that must be estimated from the data.

The simplest epidemiological model is the SI model, which only has two states (susceptible and infected). The SI model assumes that once individuals are exposed to misinformation, they shift from S to I—becoming infected and a source of spreading the misinformation themselves. Onces individuals are observed to be infected in the SI model (e.g., by reading or posting a misinformation tweet), they are always infected and never revert back to an uninfected state. Like the majority of epidemiological models, the standard SI model assumes that the "infectiousness" of a given piece of misinformation is constant throughout the population (i.e., the likelihood of sharing/receiving misinformation does not vary across nodes). Garvic and Bagdasaryan (2019), however, introduce a "fuzzy SI model" to reflect the diffusion of misinformation on twitter, with the distinctive feature of fuzziness being that it introduces a sharing coefficient (the probability of a user I successfully sharing a tweet that reaches a user S) which varies according to the I user's *influence* (which can be calculated using a range of characteristics like number of followers, expertise on the misinformation topic, etc.).

Other epidemiological models are extended to include more compartments than the simple SI binary. The SIS model, for instance, allows individuals to shift between the infected (I) and susceptible compartments—deactivating and potentially reactivating to again spread a misinformation item (Kimura et al 2009). The SIR model introduces a "recovered" state, which allows an individual—after being infected with a piece of misinformation—to become immune and not subsequently post about the misinformation item again (Maleki et al 2021 ,4).

Cheng et al (2013), for instance, develop a modified SIR model which they apply to map the diffusion of rumors on the real social network site, BlogCatalog. Their SIR model conceptualizes the R (recovered) compartment as "stiflers," or individuals who will not take part in the rumor diffusion process. As with all SIR models, transitions between compartments flow only in one direction (S → I → R), so individuals must become infected with a piece of misinformation before than can recover and become stiflers. This recovery process incorporates two distinct probabilities: (1) a probability of becoming a stifler only after encountering another infected user or another stifler; as well as (2) a probability of forgetting or disinclining to spread a rumor, which is a fixed decay parameter defined by the overall attractiveness of the rumor itself. The SIR model developed by Cheng et al (2013) is distinctive in another respect—it does not define the infectious probability as a constant in the population, but instead treats it as a variable which is a function of the strength of ties between users (susceptible users are more likely to be successfully infected by misinformation messages from other users with whom they have stronger ties).

More recently, Cho et al (2019) developed a SIR model for misinformation diffusion online which similarly includes a parameter for the *activation rate* to capture the varying probability that a given user accepts (mis)information sent from one of their contacts. The activation rate was specified as a variable reflecting the relative influence of the sending-user within the receivingusers friend group. Cho et al (2019) further introduce an *uncertainty parameter* that allow users in their model to shift back-and-forth between S, I, and R so long as: *uncertainty (regarding a piece of information) > 0*.[21]

[21] Cho et al (2013) simulate their SIR misinformation model using synthetic networks as well as a realistic network topology taken from a sample Facebook user-network (n=4039 nodes, edges=88234, average clustering coefficient=0.6055) provided by McAuley and Leskovec (2012).

The SI, SIS, and SIR models are all similiarly limited in that they only afford susceptible individuals one possibility (though sometimes a probabilistic one) as a result of newly coming in contact with an infected user: becoming infected themselves. This limitation is particularly restrictive for online social networks, where many users need time to determine whether or not they believe information sent them and may also be skeptical of the information for reasons unrelated to the strength of ties that they have with the sender.

As a result, a new epidemiological model (the SEIZ model) of online misinformation diffusion is now gaining popularity. In the SEIZ model, the population of a network is divided into four compartments: *susceptible* (S) who have not encountered the misinformation; *exposed* (E) who have encountered the misinformation and are taking a time delay to decide whether to spread it themselves; *infected* (I) who have spread the misinformation item; and *skeptical* (Z) who encountered the misinformation and made a positive decision not to tweet about it (Bettencourt et al 2006). The SEIZ epidemiological model was used by Fang et al (2013) to map the diffusion of four different news topics (true) and four rumors (false) on Twitter, with their results showing that the SEIZ model outperformed the reduced SIS model in fitting the volume of misinformation rumor tweets over time—because (E) accounts for a time delay between exposure to, and subsequent spread of, misinformation. Melaki et al (2021) use SEIZ to model the volume of misinformation tweets during the 2020 Black Lives Matter protests in Washington DC, finding that the SEIZ model was highly accurate in fitting to the spread of misinformation about this topic over time.[22]

Of course, the SEIZ model—like most epidemiological models[23]—remains limited in that it assumes that a fixed propagation parameter defines the probability that a piece of misinformation will successfully spread from an infected node to an uninfected node. This

[22] Melaki et al (2021) operationalized misinformation tweets as those including the #DCblackout hashtag—spreading a false claim that there was a communication blackout in DC because of riots.

[23] Exceptions are found in epidemiological models that treat the propagation probability as a variable that is not constant across the population, such as Cheng et al (2013) and Cho et al (2019).

parameter is referred to by Melaki et al (2021, 5) as the “basic reproduction number (R0),” which is analogous to the contagiousness of a disease: if R0 = 1, the disease is endemic because, on average, an infected person only transmits the disease to 1 other infected person; if R0 > 1, the disease is epidemic because an infected person will, on average, infect 2+ others and so the disease spreads at an exponential rate. This parameter (R0) is fixed throughout the population, as it is an attribute of the disease (or misinformation item).

Amoruso et al (2020, 852) argue that a severe weakness of epidemiological models is that they “assume that there exists a global parameter that describes the probability that a user is infected by a neighbor.” This assumption, they contend, fails to describe real life social networks in which users have wildly different probabilities of accepting (mis)information provided by their neighbors.

Topological Models

Topological models of misinformation diffusion are more flexible in that they do away with assumptions of a fixed and global contagion probability. They do so either by assuming that each node has a latent activation probability (Linear Threshold models, LT) or that each edge has a latent activation probability (Independent Cascade models, IC).

The linear threshold (LT) model is effectively a tipping-point model, in which each node is activated at given probability that is divided among all of its edges—the probability accumulates as the number of edges sending misinformation to that node increases (Wu et al 2016, 130–131). The specific activation probability varies across nodes, and in practice it is most typically assigned by selecting randomly from a normal distribution ranging from 0 to 1 (Granovetter 1978) or instead set as a fixed number (often 0.5) for all nodes within the population (Berger 2001). In the linear threshold model, the misinformation diffusion process unfolds as such: (Step 1) there is an initial set of misinformation-infected nodes and uninfected nodes; (Step 2) we newly infect every node for which the total weight of infected neighbors meets/exceeds its assigned activation threshold; (Step 3) we continue infecting until no new nodes can be infected.

The linear threshold model's most significant limitation is that it assumes users infected with misinformation never stop trying to infect their neighbors, and further, that an uninfected node is constantly receptive to activation by each of its neighbors. This is somewhat unrealistic in the social media environment, where users may give up on spreading a piece of misinformation to users who did not accept it previously—and where the effective spread of information is often between users one-to-one rather than the weight of all neighbors' opinions taken in aggregate.

To correct for these shortcomings, Kempe et al (2003) proposed the Independent Cascade (IC) model of information diffusion, which is now one of the most widely used models in studies of online misinformation. The key differences between the LT and IC models are that for the latter:

(1) the activation probability is assigned to edges, so a given node is more likely to be activated by some of its neighbors than others; and (2) nodes infected with misinformation have only one chance to infect all of their neighbors—they may not play a role in subsequent attempts to infect a neighbor with misinformation after an initial failure to do so. Otherwise, the misinformation diffusion process unfolds in similar stages to that LT model outlined above: (Stage 1) begin with an initial set of infected and uninfected nodes; (Stage 2) infected nodes make one attempt to individually infect each of their neighbors; (3) newly infected nodes subsequently make one attempt to then infect their neighbors; and (4) the process ends when no new nodes can be infected.

It is fairly common for studies seeking to identify the root sources of online misinformation to begin by assuming that the diffusion process is well-described by the Independent Cascade Model. Nguyen et al (2012), for instance, proposed an algorithm for identifying online misinformation source suspects that functioned by reversing the diffusion process of an IC model. Amoruso et al (2020) also assume that the spread of misinformation in a social network follows the IC model, and on the basis of that assumption they found that the original sources of misinformation (the *originators*) can be correctly identified at up to 80% accuracy using simulations based off of real social network structures.[24] Maryam and Ali (2019) similarly assume that misinformation diffusion follows the IC model in their study about identifying misinformation originators, doing so at a 100% accuracy level in each simulation. But unlike Amoruso et al (2020), the Maryam and Ali (2019) study assumed only a single—not multiple—originators of misinformation in a network.

Competitive Models: Information vs Misinformation

Just as one single originator of misinformation can be more consistently identified than many originators, one topic of misinformation is more readily modeled than multiple distinct topics diffusing through a network at the same time. In research on misinformation diffusion using the LT or IC models, it is often assumed that only one relevant type of (mis)information is circulating throughout a social network in order the execute the models. These models can, however, be extended to account for multiple relevant topics that diffuse in interaction with one another.

For the linear threshold (LT) appraoch, Pham et al (2020) develop a "Multiple Topics Linear Threshold" model (MT-LT) to capture the diffusion process and better train algorithms to block misinformation topics online. For the independent cascade (IC) approach, Budak et al (2011) introduce the "Multi-Campaign Independent Cascade Model" (MCICM) to model diffusion in a way that allows two competing cascades (information cascade vs misinformation cascade) to be occurring contemporaneously—which lends itself to training the development of fact-checking and misinformation correction algorithms.

These sorts of multi-topic extensions are commonly deployed to adapt theoretical models of misinformation diffusion in a way that accounts for a competitive spread with correct information (or fact-checking) within the same social network. This was explicitly the goal of the Budak et al (2011) MCICM model described above, as the authors sought to use this model to identify the

[24] The networks that Amboruso et al (2020) used for their simulations included: Epinions, Advogato, Digg Friends, Google+, Slashdot Zoo, Twitter lists, Youtube Friendship, Facebook (NIPS), and Facebook Friendships. These networks were accessed by the authors through the KONECT project database http://konect.cc/.

most influential users in a social network from which to start a counter-misinformation cascade so as to minimize the total number of nodes that ultimately adopted misinformation. Budak et al’s (2011) MCICM model largely resembles the standard Independent Cascade (IC) model discussed above. The key distinction, however, is that MCICM assumes two opposing cascades occurring at the same time. When a given node becomes active (with either correct information or misinformation), it immediately gets one chance to activate each currently inactive neighbor. Once activated, nodes never deactivate or switch their activation type (i.e., between informationand misinformation-active). The cascade continues in subsequent steps until no new node can be activated by either of the two cascades. Budak et al (2011) run simulations of the MCICM model on a sample of regional network structures obtained from Facebook (network snapshots of Santa Barbara and Monterey Bay in 2008 and 2009). The goal of these simulations was to identify optimal seed nodes for a counter-misinformation campaign.

A similar theoretical adaptation of the standard IC model called the Misinformation-Protector Independent Cascade (MPIC) model was deployed by Chen et al (2019). Their simulations, however, had a slightly different goal of minimizing the profit of edges of infected nodes—as opposed to the Budak et al (2011) goal of minimizing the total number of infected nodes at the end of the cascade.

Ghoshal et al (2019) adapt the IC model in a more unique way, notably by: (1) not imposing the assumption that an active node never shifts type (i.e., a misinformation node can be later become an active correct information node); and (2) allowing activated nodes a possibility of multiple attempts at activating their neighbors. This model of competitive information spread is referred to as the Multi-Try Independent Cascade Model (MTICM). In MTICM, a previously active node has multiple opportunities to attempt to spread its state to neighboring nodes. Moreover, Ghoshal et al (2019) assume that the probability of diffusing correct-information is always greater than the probability of diffusing misinformation (because correct information typically comes from more authoritative sources). As a result, a misinformation-activated node has a chance of accepting a correct-information activation attempt by one of its neighbors. Using this MTICM model, Ghoshal et al (2019) run a series of simulations with the objective of minimizing the total amount of time that a false rumor remains active in the network.

Moving to multi-topic Linear Threshold (LT) models, Zhang et al (2018) deploy a Separated Threshold model to describe the competitive diffusion of information and misinformation in a network. While the authors use this model to pursue twin goals of minimizing misinformation spread and maximizing correct information spread, it was originally developed to model product competition (Borodin et al 2010). In this modification of the LT model, every node has two different assigned activation thresholds, one for the acceptance of information and another for the acceptance of misinformation. Every edge is also given two separate weights, one for the impact of Node A on Node B *with respect to truth* and another *with respect to misinformation*. The LT framework begins with a given set of Truth and Misinformation Nodes, and then proceeds normally with nodes activating when the sum of their edges reach either their Truth or Misinformation threshold. Zhang et al (2018) run simulations on four real network structures (BlogCatalog, Epinions, Livemocha, LiveJournal) with the goal of evaluating an algorithm for selecting the best seed nodes from which to disseminate correct information.

In terms of epidemiological models, Tambuscio et al (2015) modify an SIS model to allow for competitive information spread by splitting the infected (I) compartment into two subcompartments: Believers (B) and Fact-Checkers (F). The transition process envisioned by the model is: Susceptible (S) → B — F → (S). The spreading function for diffusing correct/incorrect information is computed separately for each node as a function of: (1) a constant propagation parameter; (2) the (non)credibility of the piece of misinformation in question; and (3) the number of neighbors who Believe or Fact-Check. Tambuscio et al (2015) also introduce a probability of forgetting (*p-forget*) its belief in factual or non-factual information, which is how nodes transition back between B — F → S. Finally, Tambuscio et al also introduce a verifying probability (*p-verify*), which is the likelihood that each agent will fact-check the misinformation independently and so switch from B → F. The authors simulate the model on a sample Facebook network structure with the goal of identifying a threshold probability for *p-verify* that will always ensure the eventual and complete eradication of the misinformation from the network. Sulis and Tambuscio (2020) later refer to this model as the SBFC model, and they design an open source tool called NetLogo which can be utilized to run this model over any given network structure and additionally allows users to set different values for the *propagation rate, misinformation credibility, forgetting probability, and verification probability.*

2.3 Observational Work on (Mis)information Diffusion

This section discusses observational work on the diffusion of (mis)information online. It begins by discussing how observational social media data has alerted scholars to important temporal features in the diffusion of online information. Second, it highlights how the networks structures used in theoretical models may differ wildly from the actual structures governing information diffusion in real online networks. Third, it reviews empirical work exploring whether information and misinformation spread differently in online space.

Time Trends

Looking at observational data on how information spreads throughout a social network has proven valuable in identifying new parameters that need to be incorporated into conventional diffusion models, such as the Linear Threshold (LT) and Independent Cascade (IC). Saito et al (2010), for instance, observe that time stamps for real social media posts are not evenly spaced, which suggests topographical models must be updated to include a *time delay parameter* to better reflect the data generation process. Saito et al (2010) introduce a parameter accounting for an *asynchronous time delay* into both the LT and IC models, and they subsequently evaluate the performance of the Asynchronous Linear Threshold (AsLT) and the Asynchronous

Independent Cascade (AsIC) on real social network data.[25] Their results demonstrate that: (1) the AsLT and AsIC models are both, overall, highly capable of capturing the global characteristics of information diffusion in the network; and (2) that whether the AsLT or AsIC model is preferable (i.e., has a higher predictive accuracy) depends on the *topic of information* in question.

Guille and Hacid (2012) similarly apply the AsIC model to examine the dynamics of information diffusion in a real social network (Twitter). They extract a data set of 467 million Twitter posts from 20 million users for the period of June 2009–December 2009. Using this data, Guille and Hacid (2012) first show that the percentage of total-tweets issued by individual users varies in different time-blocks of the day (i.e., some users tweet more in mornings, some evenings, etc.), and further, that the probability of a successful diffusion of information between two nodes depends upon whether the sender posts information at a time when the receiver is typically active/inactive. By fitting the AsIC model to the volume of real tweets about the release date of a new iPhone over time, Guille and Hacid (2012) find: (1) that this model performs well in capturing the overall pattern of information spread—peaks and valleys in the volume of tweets over time; but (2) it does not correctly estimate the overall volume—while their AsIC gets the pattern fairly correct, it still consistently underestimates the overall volume of real tweets.

Network Structure

Many of the theoretical models of (mis)information diffusion discussed in Section 2.2 were evaluated by simulating how information spreads through real social network structures (e.g., Twitter follower links, Facebook friend links, etc.). While the results produced through this approach may be internally valid, observational work on online social networks highlights potential questions of external validity. This is because: (1) the simulations may fail to reflect how online misinformation diffuses in practice; and (2) that even real network structures of

[25] The real data utilized by Saito et al (2010) come from the Japanese blog hosting service *Doblog*, and they track the diffusion of information on a variety of topics: a "musical baton" game; a missing child; a fortune telling article; a charity site; an online flirtatious tendency test.

friends/followers online may not accurately reflect patterns in how—and with whom—social media users interact.

The first problem of simulation accuracy can be evaluated by studying whether epidemiological and topological models of information diffusion correctly describe how misinformation and fact-checking diffuse in online space. Shao et al (2018a) take on this challenge using Hoaxy, an open platform for tracking misinformation and fact-checking on Twitter which has been collecting data from the public Twitter stream since June 2016. The authors use Hoaxy to analyze the diffusion of misinformation and fact-checking on Twitter before and after the 2016 US presidential election, and their findings push against some of the optimism provided by theoretical models of how information and misinformation compete. As discussed above, Ghoshal et al (2019) assume that the probability of a node accepting correct information alway outweighs the probability of accepting misinformation. Shao et al's (2018a) empirical analysis of Twitter data, by contrast, reveals that misinformation and fact-checks separate into two distinct networks, and in the core of the misinformation network fact-checking is almost completely absent. The few fact-check Tweets that exist in this core tend to be mocking or misleading about their content. This finding suggests that correct-information cascades have a more limited utility than some theoretical models expect, and further, that the notion correct-information necessarily has a greater acceptance probability than misinformation is difficult (though perhaps not impossible) to justify in the real world. Shao et al's (2018a) findings also challenge Tambuscio et al's (2015) conclusion that misinformation can be completely eradicated from a network if the verification probability (*p-verify*) is sufficiently high—as this theoretical implication of their SBFC model appears extremely difficult to actually translate into practice.

Huberman et al (2008) provide evidence for the second external validity problem, that networks (even real ones) used in simulations may not correctly reflect how social media users link to—and interact with—one another. They do so through an observational study of real interactions between 309,740 Twitter users. The main conceptual contribution of Huberman et al (2008) is distinguishing between "*followers/followees*" (publicly declared network ties) and "*friends*" (persons with whom a user actually interacts).[26] The authors conclude that most of the declared *follower* links between nodes on Twitter were "meaningless from an interaction point of view" because a "hidden social network" of *friends* (composed of links not publicly declared) more accurately reflects how users interact and how information spreads between them. As a result, network simulations that rely on drawing edges through declared tied (i.e., official followers) may not reflect the real structure of how the network is connected in practice. For political scientists, the distinction that Huberman et al (2008) make between *following-ties* and *friendship-ties* may sound somewhat analogous to distinctions IR scholars make between *all dyads* and *politically-relevant dyads* in international conflict data.

Building upon this insight from Humberman et al (2008), Yang and Counts (2010) analyze the spread of information among Twitter users (n=3,243,437)[27] by constructing the network based on "friends" (through @username mentioning) rather than "followers." They then study how information propagates through the network with three dependent variables: (1) *Speed*, or the duration until the first diffusion of a post occurs if at all; (2) *Scale*, or the number of first-degree child nodes for a post; and (3) *Range*, or how far the diffusion chain of a post extends. Yang and Counts (2010) find that the *amount a user is mentioned* is a strong and positive predictor of *speed*, *scale*, and *range*. They further conclude that for *speed* and *range*, Tweets that appeared later in the conversation about a topic diffused faster and farther—implying that the earliest post about a topic is not necessarily the most important. For the study of misinformation, Yang and Counts' (2010) study suggests that *influential nodes* (those frequently mentioned by others) will be the best conduits for rapid and deep diffusion throughout a network.

[26] A "friend" is operationalized by Huberman et al (2008) as someone who a given user has individual directed at least two posts to using the "@username" tag.

[27] Yang and Counts' data set includes 22,242,221 posts among these users between 7/8/2009 and 8/8/2009.

Does Misinformation Diffuse Differently than Information

All of the theoretical models described in Section 2.2 are applied to the diffusion of *information* as well as *misinformation*. In addition, the discussion in this section has—until this point—focused primarily on observational research investigating the spread of *information*. This reflects a broader assumption of a shared diffusion process, one which is made explicit by Amoruso et al (2020, 849–850):

> "Even assuming that the misinformation sources have malicious intents, which may not always be the case, many users contribute to spread it further because they assume it to be true, *therefore the diffusion of true or false information should not be modeled*
> *differently*."

Of course, Amoruso et al (2020) acknowledge that the parameters in information diffusion models—whether *epidemiological* or *topological*—may need to be specified differently when misinformation is the object of interest. If, for instance, misinformation is more likely to be shared, then the "basic reproduction number"[28] (global infectiousness parameter) needs to be set higher in epidemiological models focusing on misinformation as opposed to information diffusion.

Other studies refrain altogether from assuming that information and misinformation follow similar diffusion models. Such studies instead use observational data to directly evaluate the merits of that assumption and cultivate a deeper understanding of realistic (mis)information cascades online. Zhao et al (2020) investigate whether *real news* and *fake news* propagate differently online by collecting data on social media posts from two sources: (1) a network of 973,391 Weibo users in China focusing on 1701 fake news topics and 492 real news topics; and

[28] See the above discussion of Melaki et al (2021).

(2) a network of 105,335 Twitter users in Japan posting from 3/11/2011 to 3/17/2011 when fake news about the Great East Japan earthquake was rampant. Zhao et al (2020, 1) find that "fake news spreads distinctively from real news even at early stages of propagation." First, the number of layers in the diffusion-chain of a fake news network was larger than that for real news, indicating that fake news spreads *deeper*. Second, fake news posting had a much lower fraction of repostings in the first layer than real news, but a much higher fraction of repostings in lower layers—indicating that the diffusion of fake news has distinctly a greater *scale* (i.e, number of child nodes) at lower layers of the network. Third, the authors find that real news spreads with a high degree of *heterogeneity* (meaning the network corresponds closely to a *star-like layout*) whereas fake news exhibits a much lower degree of *heterogeneity* (meaning the network deviates more heavily from a star-layout) since its "propagation involves few dominant broadcasters" (Zhao et al 2020, 7).

Vosoughi et al (2018) analyze the spread of 126,000 true and false news stories on Twitter from 2006–2017,[29] and their findings—in line with (Zhao et al 2020)—demonstrate that false news spread deeper (with a longer reshare chain) than true news. At the same time, the results from Vosoughi et al (2018) run afoul of Zhao et al's (2020) conclusion that fake news is typically propagated by a few dominant backers. Instead, Vosoughi et al (2018, 1147) observe fake news to have a high degree of *structural virality*—in the sense that it is not diffused through a single large broadcast but rather spreads through a progressive *peer-to-peer* structure over multiple generations in which no one individual is responsible for more than a small fraction of the total spread.

Friggeri et al (2014) explore how 4,761 distinct *rumors* spread on Facebook compared to *nonrumors*. The authors and identified rumors using Snopes.com, and coded rumors as being either *true* or *false*. Their analysis shows first that compared to non-rumors (general information cascades), rumors spread more *deeply* within the social network. Second, *true rumors* spread

[29] The Vosoughi et al (2018) data set includes over 4.5 million tweets from approximately 3 million different users.

more virally than *false rumors* in the sense that the former will, on average, result in a larger overall cascade. Finally, Frigerri et al (2014) demonstrate that the diffusion of rumors online (true and false) tends to be *bursty*—meaning they can lie dormant for months and spontaneously become popular again.

The uniquely bursty nature of misinformation diffusion online is corroborated by other studies. Using Twitter data, Ghenai and Mejova (2017) study the diffusion of over 13 million tweets containing health misinformation about Zika and observe bursts in volume and spaced over prolonged periods of relatively low activity. Shin et al (2018) trace the lifecyle of political rumors on Twitter over the 2012 U.S.presidential election, finding that false rumors are much more *bursty* than true rumors. The authors show that false rumors exhibit repeated peaks over an extended period of time (*burstiness)*, whereas true rumorsarewere not nearly as bursty but instead have a single prominent spike early on in their lifecycles (Shin et al 2018, 283). The study by Shin et al (2018) further indicates that when false rumors do eventually reemerge, they tend to do so in a mutated form that tilts toward evolving to become more extreme and intense over time.

Compared to true information, misinformation also exhibits unique patterns of diffusion among different types of users. Vosoughi et al (2018), for instance, observe that misinformation is significantly more likely to be spread by users who have fewer followers, follow fewer people themselves, and are overall less active on Twitter. This conclusion suggests that topological models of misinformation diffusion (e.g., LT and Ic) should account for user-characteristics when assigning probabilities that misinformation will spread from one user to another along a given edge.

Analyzing the diffusion of 209 different rumors (True and False) on Twitter, Vosoughi et al (2017) find that *true rumors* are more likely to spread from low-influence users (few followers) to high-influence users (many followers). False rumors, by contrast, are much less likely to be spread through linkages between low-to-high users. The reason, Vosoughi et al (2017, 30) contend, is likely a selection effect through which high influence users will “not risk retweeting a less known user’s information unless the person had very good reasons to believe the information is true.”

In a more recent study, Valecha et al (2021) observe that the likelihood of sharing misinformation online is affected by different factors for different types of users (i.e., *low influence* vs *high influence*). The authors study a Twitter sample of 155,589 tweets about the Zika outbreak posted between September 2015 and May 2017. Of those tweets, just under one-third (n=45,498) contained misinformation. The results from Valecha et al (2021) indicate: (1) that *temporal distance* of the health crisis has a negative effect specifically on the likelihood of *high influence users* posting a misinformation tweet—suggesting they tend to avoid “old news”; and (2) *spatial distance* between the user’s location and the crisis location has a negative effect specifically on the likelihood of *low influence users* posting a misinformation tweet—suggesting low influence users are less prone to tweet about events happening far away.

Finally, a number of studies suggest that different types of (mis)information may diffuse in unique ways, and as a result, devising one overarching model for misinformation spread writ large may be inappropriate. Vosoughi et al (2018, 1148) observe that false *political news* on Twitter: (1) diffuses deeper much more quickly than other types of false news; and (2) reached a much higher number of unique users large levels of depths in the cascade chain—particularly for depths greater than 10, in which political misinformation markedly involves more unique users that other types of misinformation.

Analyzing the spread of misinformation from 32 different public Facebook pages about conspiracy theories, Del Vicario et al (2016) offer some evidence that *conspiracy theories* diffuse differently than other types of misinformation online. In contrast to the general *bursty* misinformation diffusion pattern (Friggeri 2014; Ghenai and Mejova 2017; Shin et al 2018), *conspiracy theories* do not tend to emerge rapidly, decay and lay dormant, and then again reemerge. Instead, Del Vicario et al (2016, 555) find that conspiracy theories cascade slowly, do not decay rapidly, and gradually become propagated to more-and-more users over time.

Work on the diffusion of verifiable information (often advertisements) online indicates that individual users within a social network have different levels of influence over different topics (Li et al 2015). Based on the intuition that "a user may be influenced by her friend in some topics (e.g., sports), while remaining neutral/unaffected in others (e.g., politics)," Fan et al (2018) design a social influence analysis system ("OCTOPUS") to identify which nodes in a network are most influential for different keyword topics.[30] This system considers edge strength between two nodes to be topic dependent, an intuition which has been incorporated into theoretical models of misinformation diffusion and blocking by Pham et al (2019, 2020). Shu et al (2020) develop *FakeNewsNet*, a data repository for fake news online of two types: *political* and *entertainment*. While the Shu et al (2020) do not themselves analyze whether entertainment or political fake news diffuse in different ways, they do note that FakeNewsNet (by containing political and entertainment domains) will likely lend its self well as a tool for studying "common and different patterns for fake news under different topics."

[30] Fan et al (2018) demonstrate OCTOPUS on two different networks, one with academic citations (ACMCite) and one commercial social network for advertising services (Tencent QQ).

References

[1] Ackland, Robert and Karl Gwynn. (2021). “Truth and the Dynamics of News Diffusion on Twitter.” In *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*, edited by Rainer Greifeneder, Mariela E.Jaffe, Eryn J.Newman, and Norbert Schwarz.´ New York: Routledge.

[2] Alimardani, Mahsa, and Mona Elswah. (2020). “Online Temptations: COVID-19 and Religious Misinformation in the MENA Region.” *Social Media + Society* 6(3): 1–4.

[3] Allcott, Hunt, and Matthew Gentzkow. (2017). “Social Media and Fake News in the 2016 Election.” *Journal of Economic Perspectives* 31(2): 211–236.

[4] Allcott, Hunt, Matthew Gentzkow, and Chuan Yu. (2019). “Trends in the Diffusion of Misinformation on Social Media.” *Research & Politics* 6(2): 1–8.

[5] Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. (2020). “Evaluating the Fake News Problem at the Scale of the Information Ecosystem.” *Science Advances* 6(14): 1–6.

[6] Amoruso, Marco, Daniele Anello, Vincenzo Auletta, Raffaele Cerulli, Diodato Ferraioli, and Andrea Raiconi. (2020). “Contrasting the Spread of Misinformation in Online Social Networks.” *Journal of Artificial Intelligence Research* 69: 847–879.

[7] Ball, James. (2017). *Post-Truth: How Bullshit Conquered the World*. London: Biteback Publishing.

[8] Berger, Eli. (2001). “Dynamic Monopolies of Constant Size.” *Journal of Combinatorial Theory* 83(2): 191–200.

[9] Bessi, Alessandro, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. (2015). “Science vs Conspiracy: Collective Narratives in the Age of Misinformation.” *PloS One* 10(2).

[10] Bettencourt, Lu´ıs MA, Ariel Cintron-Arias, David I. Kaiser, and Carlos Castillo-Ch´ avez.´ (2006). “The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological Models.” *Physica A: Statistical Mechanics and its Applications* 364, 513–536.

[11] Borges, Priscila Monteiro, and Renira Rampazzo Gambarato. (2019). “The Role of Beliefs and Behavior on Facebook: A Semiotic Approach to Algorithms, Fake News, and Transmedia Journalism.” *International Journal of Communication* 13: 603–618.

[12] Borodin, Allan, Yuval Filmus, and Joel Oren. (2010). “Threshold Models for Competitive Influence in Social Networks.” *International Workshop on Internet and Network Economics*: 539–550.

[13] Budak, Ceren, Divyakant Agrawal, and Amr El Abbadi. (2011). “Limiting the Spread of Misinformation in Social Networks.” *Proceedings of the 20th International Conference on World Wide Web*: 665–674.

[14] Burel, Gregoire, Tracie Farrell, Martino Mensio, Prashant Khare, and Harith Alani. (2020).´ “Co-Spread of Misinformation and Fact-Checking Content During the COVID-19 Pandemic.” *International Conference on Social Informatics*: 28–42.

[15] Chen, Liang, Xiaohui Wang, and Tai-Quan Peng. (2018). “Nature and Diffusion of Gynecologic Cancer-Related Misinformation on Social Media: Analysis of Tweets.” *Journal of Medical Internet Research* 20(10): 1–11.

[16] Chen, Tiantian, Wenjing Liu, Qizhi Fang, Jianxiong Guo, and Ding-Zhu Du. (2019). “Minimizing Misinformation Profit in Social Networks.” *IEEE Transactions on Computational Social Systems* 6(6): 1206–1218.

[17] Cheng, Jun-Jun, Yun Liu, Bo Shen, and Wei-Guo Yuan. (2013). “An Epidemic Model of Rumor Diffusion in Online Social Networks.” *The European Physical Journal* 86(1): 1–7.

[18] Cho, Jin-Hee, Scott Rager, John O’Donovan, Sibel Adali, and Benjamin D. Horne. (2013). “Uncertainty-Based False Information Propagation in Social Networks.” *ACM Transactions on Social Computing* 2(2): 1–34.

[19] Colborne, Adrienne, and Michael Smit. (2017). “Identifying and Mitigating Risks to the Quality of Open Data in the Post-Truth Era.” *IEEE International Conference on Big Data (Big Data)*: 2588–2594.

[20] Del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. (2016). “The Spreading of Misinformation Online.” *Proceedings of the National Academy of Sciences* 113(3): 554–559.

[21] Dickey, Lauren. (2019). “Confronting the Challenge of Online Disinformation in Taiwan.” *Disinformation, Cybersecurity, & Energy Challenges*.

[22] Fan, Ju, Jiarong Qiu, Yuchen Li, Qingfei Meng, Dongxiang Zhang, Guoliang Li, Kian-Lee Tan, and Xiaoyong Du. (2018). “OCTOPUS: An Online Topic-Aware Influence Analysis System for Social Networks.” *IEEE 34th International Conference on Data Engineering (ICDE)*: 1569–1572.

[23] Jin, Fang, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. (2013). "Epidemiological Modeling of News and Rumors on Twitter." *Proceedings of the 7th Workshop on Social Network Mining and Analysis* 1–9.

[24] Fletcher, Richard, Alessio Cornia, Lucas Graves, and Rasmus Kleis Nielsen. (2018). "Measuring the Reach of 'Fake News' and Online Disinformation in Europe." *Australasian Policing* 10(2): 1–10.

[25] Friggeri, Adrien, Lada Adamic, Dean Eckles, and Justin Cheng. (2015). "Rumor Cascades." *Proceedings of the International AAAI Conference on Web and Social Media* 8(1).

[26] Gavric, Dubravka, and Armen Bagdasaryan. (2019). "A Fuzzy Model for Combating Misinformation in Social Network Twitter." *Journal of Physics: Conference Series* 1391(1).

[27] Amira Ghenai and Yelena Mejova. (2017). "Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter." *arXiv preprint*.

[28] Ghoshal, Arnab Kumar, Nabanita Das, and Soham Das. (2019). "Misinformation Containment in OSNs Leveraging Community Structure." *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*: 1–6.

[29] Gorrell, Genevieve, Mehmet E. Bakir, Ian Roberts, Mark A. Greenwood, Benedetta Iavarone, and Kalina Bontcheva. (2019). "Partisanship, Propaganda, and Post-Truth Politics: Quantifying Impact in Online Debates." *ArXiv preprint*.

[30] Granovetter, Mark. (1978). "Threshold Models of Collective Behavior." *American Journal of Sociology* 83(6): 1420–1443.

[31] Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. (2019). "Fake News on Twitter during the 2016 US Presidential Election." *Science* 363(6425): 374–378.

[32] Guess, Andrew, Brendan Nyhan, and Jason Reifler. (2018). "Selective Exposure to Misinformation: Evidence from the Consumption of Fake News During the 2016 US Presidential Campaign." *European Research Council* 9(3).

[33] Guess, Andrew, Jonathan Nagler, and Joshua Tucker. (2019). "Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5(1): 1–8.

[34] Guille, Adrien, and Hakim Hacid. (2012). "A Predictive Model for the Temporal Dynamics of Information Diffusion in Online Social Networks." *Proceedings of the 21st International Conference on World Wide Web*: 1145–1152.

[35] Ha, Louisa, Loarre Andreu Perez, and Rik Ray. (2021). “Mapping Recent Development in Scholarship on Fake News and Misinformation, 2008 to 2017: Disciplinary Contribution, Topics, and Impact.” *American Behavioral Scientist* 65(2): 290–315.

[36] Huberman, Bernardo A., Daniel M. Romero, and Fang Wu. (2008). “Social Networks That Matter: Twitter under the Microscope.” *arXiv preprint*: 1–9.

[37] Hussein, Eslam, Prerna Juneja, and Tanushree Mitra. (2020). “Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube.” *Proceedings of the ACM on HumanComputer Interaction* 4: 1–27.

[38] Ismailova, Larisa, Viacheslav Wolfengagen, Sergey Kosikov, Mikhail Maslov, and Juliane Dohrn. (2020). “Semantic Models to Indicate Post-Truth with Fake News Channels.” *Procedia Computer Science* 169: 297–303.

[39] Li, Yuchen, Dongxiang Zhang, and Kian-Lee Tan. (2015). “Real-Time Targeted Influence Maximization for Online Advertisements.” *Proceedings of the VLDB Endowment* 8(10): 1070–1081.

[40] Kempe, David, Jon Kleinberg, and Eva Tardos. (2003). “Maximizing the Spread of Influence´ through a Social Network.” *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 137–146.

[41] Kimura, Masahiro, Kazumi Saito, and Hiroshi Motoda. (2009). “Efficient Estimation of Influence Functions for SIS model on Social Networks.” *Twenty-First International Joint Conference on Artificial Intelligence*.

[42] King, Gary, Jennifer Pan, and Margaret E. Roberts. (2017). “How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument.” *American Political Science Review* 111(3): 484–501.

[43] Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. (2000). “Misinformation and the Currency of Democratic Citizenship.” *The Journal of Politics* 62(3): 790–816.

[44] Lewandowsky, Stephan, Werner GK Stritzke, Klaus Oberauer, and Michael Morales. (2005). “Sectionry for Fact, Fiction, and Misinformation: The Iraq War 2003.” *Psychological Science* 16(3): 190–195.

[45] Maleki, Maryam, Esther Mead, Mohammad Arani, and Nitin Agarwal. (2021). “Using an Epidemiological Model to Study the Spread of Misinformation during the Black Lives Matter Movement.” *ArXiv preprint*.

[46] Maryam, Amrah, and Rashid Ali. (2019). “Misinformation Source Identification in an Online Social Network.” *2019 IEEE 5th International Conference for Convergence in Technology*: 1–5.

[47] McAuley, Julian J., and Jure Leskovec. (2012). “Learning to Discover Social Circles in Ego Networks.” *Neural Information Processing Systems*.

[48] Mocanu, Delia, Luca Rossi, Qian Zhang, Marton Karsai, and Walter Quattrociocchi. (2015). “Collective Attention in the Age of (Mis)information.” *Computers in Human Behavior* 51: 1198–1204.

[49] Myers, Steven Lee, and Paul Mozur. (2019). “China is Waging a Disinformation War against Hong Kong Protesters.” *The New York Times*. https://www.nytimes.com/2019/08/13/world/asia/hong-kong-protests-china.html.

[50] Nguyen, Dung T., Nam P. Nguyen, and My T. Thai. (2012). “Sources of Misinformation in Online Social Networks: Who to Suspect?” *IEEE Military Communications Conferenc*: 1–6.

[51] Pham, Dung V., Hieu V. Duong, Canh V. Pham, Bui Q. Bao, and Anh V. Nguyen. (2019). “Multiple Topics Misinformation blocking in Online Social Networks.” *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*: 1–6.

[52] Pham, Dung V., Giang L. Nguyen, Tu N. Nguyen, Canh V. Pham, and Anh V. Nguyen. (2020). “Multi-Topic Misinformation Blocking with Budget Constraint on Online Social Networks.” *IEEE Access* 8: 78879–78889.

[53] Popat, Kashyap, Subhabrata Mukherjee, Jannik Strotgen, and Gerhard Weikum. (2018).¨ “‘CredEye: A Credibility Lens for Analyzing and Explaining Misinformation.” *Companion Proceedings of the The Web Conference*: 155–158.

[54] Qiu, Xiaoyan, Diego FM Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. (2017). “Limited Individual Attention and Online Virality of Low-Quality Information.” *Nature and Human Behavior* 1(7): 1–7.

[55] Roozenbeek, Jon, and Sander Van Der Linden. (2019). “The Fake News Game: Actively Inoculating Against the Risk of Misinformation.” *Journal of Risk Research* 22(5): 570–580.

[56] Saito, Kazumi, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. (2010). “Selecting Information Diffusion Models of Social Networks for Behavioral Analysis.” *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*: 180–195.

[57] Schwarz, Norbert and Madeline Jalbert. (2021). “When (Fake) News Feels True: Intuitions of Truth and the Acceptance and Correction of Misinformation.” In *The Psychology of Fake*

News: Accepting, Sharing, and Correcting Misinformation, edited by Rainer Greifeneder, Mariela E.Jaffe, Eryn J.Newman, and Norbert Schwarz. New York: Routledge.´

[58] Shahi, Gautam Kishore, Anne Dirkson, and Tim A. Majchrzak. (2021). "An Exploratory Study of COVID-19 Misinformation on Twitter." *Online Social Networks and Media* 22.

[59] Shao, Chengcheng, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. (2018a). "Anatomy of an Online Misinformation Network." *PloS One* 13(4).

[60] Shao, Chengcheng, Pik-Mai Hui, Pengshuai Cui, Xinwen Jiang, and Yuxing Peng. (2018b). "Tracking and Characterizing the Competition of Fact Checking and Misinformation: Case Studies." *IEEE Accesss* 6: 75327–75341.

[61] Shin, Jieun, Lian Jian, Kevin Driscoll, and Franc¸ois Bar. (2018). "The Diffusion of Misinformation on Social Media: Temporal Pattern, Message, and Source." *Computers in Human Behavior* 83: 278–287.

[62] Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. (2020). "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information fro Studying Fake News on Social Media." *Big Data* 8(3): 171–188.

[63] Sommariva, Silvia, Cheryl Vamos, Alexios Mantzarlis, Lillie Uyen-Loan Dˆ ao, and Dinorah` Martinez Tyson. (2018). "Spreading the (Fake) News: Exploring Health Messages on Social Media and the Implications for Health Professionals Using a Case Study." *American Journal of Health Education* 49(4): 246–255.

[64] Suarez-Lledo, Victor, and Javier Alvarez-Galvez. (2021). "Prevalence of Health Misinformation on Social Media: Systematic Review." *Journal of Medical Internet Research* 23(1).

[65] Sulis, Emilio, and Marcella Tambuscio. (2020). "Simulation of Misinformation Spreading Processes in Social Networks: An Application with NetLogo." *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*: 614–618.

[66] Tambuscio, Marcella, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. (2015). "Fact-Checking Effect on Vial Hoaxes: A Model of Misinformation Spread in Social Networks." *Proceedings of the 24th international conference on World Wide Web*: 977–982.

[67] Treen, Kathie M. d'I., Hywel TP Williams, and Saffron J. O'Neill. (2020). "Online Misinformation about Climate Change." *Wiley Interdisciplinary Reviews: Climate Change* 11(5): 1–20.

[68] Van der Linden, Sander, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. (2017). "Inoculating the Public against Misinformation about Climate Change." *Global Challenges* 1(2): 1–7.

[69] Vosoughi, Soroush, Mostafa 'Neo Mohsenvand, and Deb Roy. (2017). "Rumor Gauge: Predicting the Veracity of Rumors on Twitter." *ACM Transactions on Knowledge Discovery from Data* 11(4): 1–36.

[70] Vosoughi, Soroush, Deb Roy, and Sinan Aral. (2018). "The Spread of True and False News Online." *Science* 359(6380): 1146–1151.

[71] Wang, Nan, Li Yu, Ni Ding, and Dong Yang. (2014). "Containment of Misinformation Propagation in Online Social Networks with given Deadline." *Pacific Asia Conference on Information Systems*: 46–54.

[72] Wang, Yuxi, Martin McKee, Aleksandra Torbica, and David Stuckler. (2019). "Systematic Literature Review on the Spread of Health-Realted Misinformation on Social Media." *Social Science & Medicine* 240: 1–12.

[73] Waisbord, Silvio. (2018). "The Elective Affinity between Post-Truth Communication and Populist Politics." *Communication Research and Practice* 4(1): 17–34.

[74] Weng, Lilian, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. (2012). "Competition Among Memes in a World with Limited Attention." *Scientific Reports* 2(1): 1–9.

[75] Wu, Liang, Fred Morstatter, Xia Hu, and Huan Liu. (2016). "Mining Misinformation in Social Media." In *Big Data in Complex and Social Networks*, edited by My T.Thai, Weili Wu, Hui Xiong. Boca Raton: CRC Press.

[76] Yang, Jiang, and Scott Counts. (2010). "Predicting the Speed, Scale, and Range of Information Diffusion on Twitter." *Fourth International AAAI Conference on Weblogs and Social Media*: 355–358.

[77] Zhang, Huiling, Alan Kuhnle, J. David Smith, and My T. Thai. (2018). "Fight under Uncertainty: Restraining Misinformation and Pushing out the Truth." *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*: 266–273.

[78] Zhao, Zilong, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. (2020). "Fake News Propagates Differently from Real News Even at Early Stages of Spreading." *EPJ Data Science* 9(1): 1–14.

[79] Zubiaga, Arkaitz, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. (2016). “Analyzing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads.” *PloS One* 11(3).

3. Foreign Influence and Online Misinformation

Prepared primarily by SS

This section addresses the topic of foreign influence operations that disseminate online misinformation. The first section reviews general strategies and objectives of state-sponsored misinformation campaigns. The second section then discusses work that analyzes the online behavior of state trolls and misinformation agents, while the third section proceeds efforts to systematically identify online state propaganda and misinformation purveyors. The fourth section begins a shift from reviewing misinformation actors and strategies to covering the outcomes of foreign misinformation campaigns. This section emphasizes research that investigates whether—and how much—foreign misinformation campaigns are effective as well as what sorts of influence these campaigns have over public opinion and activity in target states.

3.1 Strategies and Objectives of Foreign Misinformation

Bradshaw and Howard (2017) provide a broad and cross-national view of how state governments engage in social media manipulation using different types of "cyber troops." The authors (2017, 2) define "cyber troops" as including "government, military, or political party teams committed to manipulating public opinion over social media."[1] Based on an analysis of social media operations in 28 countries[2] (their budgets, staffing, organizational behavior, communication strategies), Bradshaw and Howard (2017) give a useful starting point for assessing the varying strategies that state-sponsored "cyber troops" deploy through social media: (1) United Kingdom

[1] Cyber troops may also include *private contractors* who are hired by a government or *volunteers* who select into collaboration with the government and spread state-sponsored narratives online.

[2] Argentina, Azerbaijan, Australia, Bahrain, Brazil, China, the Czech Republic, Ecuador, Germany, India, Iran, Israel, Mexico, North Korea, the Philippines, Poland, Russia, Saudi Arabia, Serbia, South Korea, Syria, Taiwan, Turkey, Ukraine, the United Kingdom, the United States, Venezuela and Vietnam.

cyber troops spread dynamic narratives to combat terrorist propaganda online; (2) Polish cyber troops individually target influencers—prominent bloggers, journalists, activists—to convince them and their followers to hold certain beliefs which, as Gorwa (2017)[3] notes tend to be disproportionately right-wing and pro-PiS; (3) Saudi Arabian cyber troops spam trending Twitter hashtags to wash out domestic criticism or undesired online speech with a wave of unrelated content; (4) Turkish and Russian cyber troops advance targeted smear campaigns against domestic activists and critics.

Martin and Shapiro (2019) analyze 53 Foreign Influence Efforts (FIEs)[4] targeting 24 different countries between 2013 and 2018. The authors conceive of the potential objectives associated with FIEs quite generally, allowing them to range from seeking to influence elections, encourage polarization, support right-wing movements, discredit target state institutions, undermine target state leaders, or shift the policy agenda. The vast majority of the campaigns identified by Martin and Shapiro (2019, 3) were conducted by Russia (72%), with the remainder being split primarily between China, Iran, and Saudi Arabia. While FIE tactics varied, the vast majority (90%) used trolls, over half used automated methods of spreading a particular message, and over half also used fake social media accounts. Trolling, bots, and hashtag hijacking tactics are most commonly used in unison, and thus may be best treated as a joint strategy (rather than set of discrete strategies). Finally, while Martin and Shapiro (2019, 10) find that Russian FIEs most typically promote the interests of right-wing parties (e.g., Alternative for Germany, Lega Nord), there are rarer cases where Russian FIEs support left-wing parties as well (e.g., left-leaning parties in Spain pushing for Catalonia independence).

[3] Gorwa (2017) offers a more detailed analysis of Polish social media and the role of bots, trolls, and fake news.

[4] FIEs are defined as (1) coordinated campaigns by states aiming to impact politics in target states, (2) which occur through media channels—including social media, and (3) use content that is designed to look indigenous in the target state (Martin and Shapiro 2019, 1).

Russian Objectives and Strategies

A significant amount of research on online foreign influence operations focuses on activities specifically sponsored by the Russian government. This is for three reasons. First, Russian agents are highly active in online misinformation campaigns. Second, the topic is politically salient to American audiences due to Russian efforts to interfere in US elections. Third, there is a large amount of data accessible on Russian-sponsored misinformation online, much of which was made public by the US House Intelligence Committee—including social media accounts affiliated with Russia's Internet Research Agency (IRA) as well as IRA ad buys on Facebook. Helmus et al (2018) assess the objectives and strategies of Russia's social media propaganda in Eastern Europe. They conclude that one of the overarching goals of Russia's online misinformation operations is to intensify political polarization in target states by driving "wedges between ethnic Russian... populations... and their host governments" (Helmus et al 2018, x). The authors further assess the goals of Russian social media propaganda in the West, identifying additional goals of these activities in the form of achieving policy paralysis by "sowing confusion, stoking fears, and eroding trust in... democratic institutions" (Helmus et al 2018, x). Moving beyond the foreign objectives of Russia's online influence operations, Pomerantsev (2015, 42) observes that Russian state propaganda almost always has a dual (foreign & domestic) purpose—cultivating political cynicism among the Russian public by showing how easily democracy is manipulated and how any political alternatives to the authoritarian status quo are just as, if not more, corruptible.

Proceeding to an empirical analysis of a Russian language Twitter database geo-referenced to accounts in Estonia, Latvia, Lithuania, Moldova, Belarus, and Ukraine, Helmus et al (2018) identify an online community of approximately 40,000 pro-Russian propagandists active in these countries who focused heavily on spreading a pro-Russian view of the Ukraine-Russia conflict.[5]To spread their pro-Russia message online, propagandists active on Eastern European social media

[5] The Twitter network constructed by Helmus et al (2018) contains n=424,979 nodes (users).

engage in a variety of activities including news tweets, non-attributed comments on web pages, troll and bot accounts, and fake hashtag and Twitter campaigns (Helmus et al 2018, ix).

Hanlon (2018, 1)[6] argues that Russian online influence operations pursue the underlying objectives of amplifying polarization in target states, undermining their democratic institutions, enhancing internal political divisions, and weakening Western international alliances. Hanson's (2018) review of Russian online influence operations disaggregated activities deployed on distinct social media platforms. On Facebook, IRA ad buys focus heavily on promoting divisive political issues while trolls generally avoid manufacturing new disputes between groups but instead insert themselves into—and amplify—already existing issues that are socially and politically divisive (Hanson 2018, 4). On Instagram, IRA foreign influence activities similarly aim to boost politically politically divisive topics. On Twitter, Russian operatives mask themselves as local actors/officials to gain credibility—whether by pretending to be US citizens, veterans, party officials, or news organizations. Forum based sites like Reddit and 4Chan are particularly useful for "laundering" misinformation across the broader social media space because of their anonymous nature. Pro-Russian trolls can post or promote divisive content on these forums, which can later be disseminated to different social media sites in a way that affords the content more credibility but also distances Russian actors from the narrative's origination.

Moving beyond social media websites, Orttung and Nelson (2019) examines the Russian government's strategies for spreading state-supported narratives online through RT (formerly Russia Today) programming on YouTube. Their study analyzes a dataset of 70,220 YouTube videos posted by RT between February 2015–January 2017, and it argues that Russian-sponsored media heavily prioritizes online YouTube programming as a "relatively inexpensive way of gaining audience share when accessing traditional television was expensive" (77). RT's channels on YouTube are divided linguistically and regionally (English, Arabic, Spanish, German, French,

[6] Hanson (2018) reviews research on Russian social media operations, but does not collect any original data or conduct an independent empirical analysis.

Chinese, Russian). Orttung and Nelson (2019, 82) conclude that RT most heavily prioritizes videos targeting Arab audiences (25.8% of all videos in their sample), with Russian audiences being second (19.4% of all videos), and Spanish and English audience being third (tied at 11% of all videos). The emphasis on Middle East audiences most typically reflected an objective of spreading pro-Russian information on the war in Syria (with "Syria" being the 2nd most frequent word used in video titles, following only "Russia"). Across the board, a common objective in RT's YouTube messaging is to "promote a positive view of Russia and its foreign policy objectives abroad" (Orttung and Nelson 2019, 83). At the same time, this objective was not fully achieved—Arabic audiences, for instance, proved difficult to reach and turn into viewers despite RT prioritizing them with its programming (89).

Since the 2016 US presidential election, research on online misinformation has grown highly attentive to Russian operations seeking to influence elections in Western democracies. Hansen and Lim (2019, 151) identify three main types of activities that Russian actors use to influence voters overseas: (1) *doxing*, or stealing and publicly releasing confidential information about individuals/organizations; (2) *fake news*, or publishing and promoting misleading content in online social networks; and (3) *trolling*, or inundating social media platforms with provocative and inflammatory posts. McFaul and Kass (2019, 3) contend that the joint use of these tactics to influence the 2016 US election reflected Vladimir Putin's interests in weakening Hillary Clinton's candidacy, supporting Donald Trump, delegitimizing the American electoral process, and sowing division within the US electorate.

In 2017, Facebook openly acknowledged that 470 inauthentic accounts based in Russia purchased over 3,000 ads on its platform in the lead-up to the 2016 election (Stamos 2017).[7] Dutt et al (2018) analyze all of these ads[8] to determine which types were the most and least effective. While the topic of effectiveness will be covered in the fourth section of this section, Dutt et al (2018, 1) conclude that the emphasis on ads with negative sentiment which focused on polarizing

[7] Alex Stamos was Chief Security Officer for Facebook.

[8] The ads were released by the House Intelligence Committee in May 2018.

events that occurred in the past suggest Russia's 2016 Facebook ad campaign primarily desired "to sow division rather than sway the election." Of course, these two goals are not necessarily mutually exclusive. Linvill et al's (2019) analysis of tweets from IRA-linked accounts in the month preceding the 2016 election found that the bulk of Russia's election operations on Twitter supported right-wing ideas and candidates, specifically being pro-Donald Trump. As a result, Linvill et al (2019, 298) argue that the IRA's foreign influence operation simultaneously aimed to sow discord and support Trump's candidacy.

Kim et al (2018) develop a real-time digital ad tracking tool to analyze 5 million paid ads on Facebook between September 28, 2016 and November 8, 2016. Their study uncovers a large number of groups that bought ads promoting divisive issues related to the 2016 U.S. election campaign, but which also failed to file reports with the Federal Election Commission. One-sixth of all of these suspicious groups turned out to be based in Russia, and the ads they disseminated revolved around polarizing issues in US politics, such as: guns, LGBTQ rights, immigration, and race. This online election influence operation was not limited to Facebook, as Kremlin-linked actors purchased similarly divisive ads on Google and Twitter (Romm and Wagner 2017).

Terrorist Objectives and Strategies

State governments are not the only actors that pursue online foreign influence operations. At the sub-state level, terrorist organizations (most notably the Islamic State) are also active in spreading propaganda to foreign audiences using social media. The primary objective behind these operations is to promote and expand recruitment, particularly among disaffected and isolated Western youths. Online propaganda by ISIS invokes themes of offering purpose and camaraderie, and at its height the organization depicted the territories that it held as a utopian society (Gertsel 2016, 1). ISIS propaganda further seeks to cultivate a sense of urgency and imminent victory to better attract potential recruits (Fernandez 2015). And while the US media tends to focus on ISIS brutality, violence and barbarism, the groups online propaganda incorporates strategies to cultivate support and sympathy among foreign audiences: (1) mercy; (2) victimhood—by emphasizing collateral damage caused by the enemy; (3) military successes; (4) belonging; and (5) utopianism of the Caliphate (Aly et al 2017, 4). Rudner (2017) identifies additional online propaganda objectives beyond recruitment in his analysis of Al-Qaeda's internet activity—ideological indoctrination, soliciting funds, providing sympathizers with training in weapons, explosives, and tactics.

To amplify their message to a broader online audience, ISIS social media operators use a Twitter app (called "The Dawn of Glad Tidings") to spread mass posts using registered users' Twitter accounts. Farwell (2014, 51), for instance, observes that app members posted 44,000 tweets per day as ISIS marched into Mosul. Moreover, to reach larger segments of the online population, ISIS routinely hijacked already trending hashtags (e.g., #Brazil2014 during the World Cup) and included them in its own tweets to facilitate broader circulation (Farwell 2014).

While ISIS conveys a sense that its online postings are organic and authentic messages by rank-and-file group members, the reality is a that the organization's online propaganda activities are highly centralized. Klausen (2015) analyzes Twitter posts by 59 Western-origin ISIS members between January and March 2014 to assess the degree to which their postings were either organic or officially controlled. Her results show that the majority of these accounts focused on spreading posts provided from more central feeder accounts, and only rarely posted original content. Klausen et al (2012) further assess the activity of 41 Western jihadist accounts on YouTube affiliated with Al-Qaeda and find a high degree of centralization and coordination between them. These YouTube accounts also engaged in a high level of reposting each other's content in order to build in a degree of redundancy—allowing content to persist online even if individual accounts were taken down.

3.2 Online Behavior of Foreign Misinformation Agents

How do state-sponsored trolls and misinformation agents operate online, what types of content do they produce, and what sorts of issues do they seek to amplify on social media? This section reviews research addressing these questions, the bulk of which utilizes newly public data on social media accounts and posts affiliated with Russia's Internet Research Agency (IRA).

Account Activity and Information

Zannettou et al (2019a) examine the online behavior of Russian trolls by analyzing a subset of 27,000 tweets posted by 1,000 different users with know ties to the IRA. To identify distinctive behaviors associated with foreign influence trolling, the authors compare these tweets with those of a randomly drawn set of Twitter users. In terms of content, Zannettou et al (2019a) find that Russian trolls tend to tweet about highly specific world events that are salient in the news cycle (e.g., Charlottesville protests, ISIS operations). Unlike the general Twitter population, most IRA associated accounts made posts from Twitter's web client (as opposed to the mobile app), and they were registered primarily in a few concentrated countries (USA, Germany, Russia)—likely because these are their primary target populations and appearing to be local enhances account credibility. Moreover, the IRA accounts had a tendency to "reset" themselves intermittently over time—deleting old tweets and changing profile information to take on a new identity (Zannettou 2019a, 219).

Human troll users are not the only ones that mask or shift their identities over time. Hegelich and Janetzko's (2016) analyze 1,740 social bots that were active in publishing information about the Ukraine-Russia conflict on Twitter, and find that bots use mimicry tactics to avoid detection and to attract a broader audience. While the bots are very much involved in posting about political topics and boosting polarizing hashtags (a majority of their online activity), they also camouflage themselves by making inane posts about movies, memes, and jokes—a strategy that makes them harder to identify and also makes them look more "normal" to human users (Hegelich and Janetzko 2016, 582).

While these activities give an impression that Russia influence operations are highly capable of avoiding detection, we should be cautious against overstating the degree of sophistication in this regard. As an example, Facebook ads launched by the IRA were heavily clustered around being released between 9:00am–6:00pm MST (standard business hours in St. Petersburg), which suggests that—at least for online ads purchased during the 2016 US election cycle—few serious attempts were made to obfuscate the point of origin being in Russia (Boyd et al 2018, 2).

Strategically Timing Influence Operations

Howard et al (2019) analyze the IRA's online activity using data (on accounts, posts, ad buys) provided by Facebook, Twitter, and Google. Their analysis finds IRA activity initially focused primarily on Twitter in 2013, but quickly expanded to include broader operations on Facebook, Instagram and YouTube. The frequency of IRA posts on social media also exhibits a clear cut temporal trend that maps onto important dates in the US political calendar, crises, and international events—most likely due to the political nature of foreign influence objectives discussed above. This temporal pattern is corroborated by Zannettou et al (2019a) who observe a peak in IRA activity on Twitter just before the second US presidential debate (10/9/2016). Howard et al's (2019, 3) analysis reinforces the notion that IRA influence operations online pursue an overarching goal of inciting political polarization and discrediting political institutions in the United States— with a large number of posts campaigning for black Americans to boycott elections, for Mexican Americans to distrust US institutions, encouraging right-wing extremism, and spreading political conspiracies and misinformation.

Boyd et al (2018) analyze IRA Facebook ads before and after the 2016 US election, and find that the specific topics used to incite polarization varied over time. While ads related to civil rights generally emerged long before the election and persisted after, those on politically polarizing current events like "police brutality" and ”Black Lives Matter" scaled up rapidly right before election day and virtually disappeared after the election (Boyd et al 2018, 4). This suggests that IRA polarization efforts were attuned to influencing whether and how Americans voted in the election, and further, that some topics are treated as more effective vehicles for promoting polarization and cynicism than others. Etudo et al (2019) find that US elections are not the only noteworthy events around which IRA ad buys are strategically timed. In their analysis of Facebook ads purchased by the IRA between 2015–2017, the authors find that larger ad buys related to police brutality are carefully timed to coincide with periods of higher unrest in the United States, and further, that ads published by the IRA are significantly more inflammatory during periods of unrest than during periods of relative calm (Etudo et al 2019, 899).

Amplifying Both Sides of Polarizing Issues

Even for highly divisive and salient topics in the US, Russian online trolls are sometimes able to acquire a significant voice on both sides of the political conversation online. Stewart et al (2018), for instance, show that Russian IRA-linked Twitter accounts were able to reach large audiences on disparate sides of the Black Lives Matter movement between December 2015 and October 2016. Based on an analysis of 58,812,322 tweets with shooting related keywords, Stewart et al show that the online conversation broadly separated into two distinct clusters: (1) a right cluster, associated with #MAGA, #2ndAmendment, and pro-Trump posts; and (2) a left cluster, associated with #imwithher, #feelthebern, and #blacklivesmatter posts. Russian IRA accounts were active in both of these clusters. And though they were generally more successful in infiltrating the leftleaning cluster, Russian trolls were in the top-percentile of users in retweet count in both the left and right clusters of online discourse on the topic.

Even before the polarization of US COVID-19 vaccine discourse, Russian trolls were active in promoting polarization and misinformation about public health topics online. Broniatowski et al (2018) scrape a 1% sample of all tweets from Twitters API in addition to a sample of tweets containing vaccine related keywords between 2014–2017. Their results demonstrate that Russian trolls post about vaccines at a significantly higher rate than average Twitter users, and further, that the content of Russian troll tweets is fairly balanced between the pro- and anti-vaccination camps. This balanced-strategy, the authors conclude, is broadly consistent with an objective of promoting polarization and discord rather than any single favored policy view (Broniatowski et al 2018, 1379). Walter et al (2020) build upon this research by analyzing vaccine related content in a sample of 2.82 million tweets published by IRA-affiliated accounts between 2015–2017. Their study indicates an association between local partisan personas adopted by IRA accounts (i.e., masking as pro-Trump or anti-Trump Americans) and vaccine related online content. While pro-Trump IRA accounts expressed high levels of anti-vaccine sentiment online, anti-Trump IRA accounts did the opposite (Walter et al 2020, 718)—which may have the intended effect of amplifying partisan polarization around public health topics.

Spangher et al (2018) show that the IRA's efforts to incite ideological polarization in the US are heterogeneous across different social media platforms. Using data disclosed by the House Intelligence Committee on IRA Facebook ads (n=3,393) and Twitter handles associated with 2.9 million Tweets, the authors conclude that on Facebook the IRA devoted more resources toward promoting left-leaning content than it did toward right-leaning content. By contrast, IRA content on Twitter was more balanced overall, though right-leaning content still received more traffic. Spangher et al (2018) further find that IRA accounts frequently posted about new and rapidly evolving local news stories, which they associate with a strategy for capturing clicks from users interested in the topic by posting before local news agencies could cover the stories themselves. Indeed, Farkas and Bastos' (2018) analysis of 4,539 tweets by IRA accounts between 2012–2017 shows that the "lion's share of [the IRA's] work focus[es] on US daily news activity and the diffusion of polarized news." To this end, Farkas and Bastos (2018, 284–285) find that IRA accounts strategically adopt identities that mimic local news outlets and also systematically tweet more about news topics that are contentious and amplify concerns related to public security.

While IRA trolls simultaneously promote right-leaning and left-leaning content to further drive polarization, we should not conclude that foreign influence operations are generally non-partisan (or bi-partisan). Zannettou et al (2019b) compare 10 million posts on Twitter and Reddit by users identified as being affiliated with the Russian and Iranian governments. Their analysis covers a time period of 2012–2018, and it shows clear and diverging partisan preferences between Russian and Iranian state-sponsored trolls; the former were significantly more pro-Trump in their online activity, while the latter were overwhelmingly anti-Trump.

3.3 Identifying State-Sponsored Misinformation

Following Russian efforts to use social media to influence Americans' attitudes and behavior around the 2016 election, a growing body of work has begun to focus specifically on how to identify foreign propaganda online. Barron-Cedeno et al (2019) devised an algorithm that automates the detection of online propaganda—not just foreign-sponsored propaganda—called *Proppy*. The algorithm is trained on a dataset of known propaganda sources to detect common stylistic features, and it is able to be used for real-time monitoring of online news, organizing articles, and outputting an estimated likelihood that any individual article contains propagandistic content.

While the *Proppy* system is oriented toward analyzing individual news articles, alternative identification methods are focused more broadly at detecting whether particular Twitter accounts are state-sponsored trolls. Im et al (2020) use machine learning models to identify whether specific Twitter users are IRA affiliated trolls by discriminating between a sample of 2,200 known IRA Twitter handles and a control sample of 170,000 control accounts. Their model achieves a precision level of 78.5%, and relies chiefly upon classifying accounts according to profile information, posting behavior, linguistic features, and word usage. On average, accounts flagged as being IRA affiliated had: (1) more follows and followers than unflagged accounts; (2) were more recently created; (3) use hashtags with a higher frequency; (4) had a higher rate of retweeting political content; and (5) had more tweets in Russian (Im et al 2020, 8).

Boyd et al's (2018) analysis of IRA posts during the 2016 election cycle gives additional reasons for optimism about the prospects of correctly classifying Russian propagandistic content online. Their analysis of the linguistic features of IRA Tweets in English suggests clear differences in semantic content between IRA trolls and native English speakers.

While efforts to identify foreign actors spreading misinformation and propaganda online are moving forward, Zannettou et al's (2019b) analysis of 10 million Twitter and Reddit posts by Russian and Iranian trolls suggests the process will be highly difficult and ultimately imperfect.

This is primarily because the accounts in their sample do not exhibit a consistent behavior over time, but instead change their behavior is somewhat unpredictable ways. As a result, automated detection efforts that prove effective at one point in time may be wholly incapable of correctly classifying future online foreign influence campaigns that adopt different posting patterns and strategies.

Detection methods are also being used to identify whether online content is propaganda from terrorist organizations. Tundis et al (2020) devise a conceptual "pyramid model" that aims to assess whether—and how much—individual users are radicalized in support of terrorist organizations. Their conceive of three emotional and sentimental dimensions that are useful for flagging radicalization in online content: (1) emotion—sadness, joy, anger, fear, trust, disgust; (2) psychological affinity—polarity, attention, and sensitivity to groups or issues; and (3) semantic context—the use of persuasion or coercion. The authors then use these dimensions to classify levels of extremism (organized in a pyramid of high-to-low radicalization) in a sample of 17,000 Twitter posts from 100 pro-ISIS users. They are explicit in noting that their sample selects on the dependent variable, and therefore, does not lend itself to making effective classifications between "terrorist propaganda" and "non-terrorist propaganda" (Tundis et al 2020, 4).

In an effort to move beyond this limitation, Mussiraliyeva et al (2020) use machine learning methods to facilitate the automated detection of extremist content online. They scrape posts from the Vkontakte social network, which is popular in post-Soviet countries (namely the Commonwealth of Independent States), and then manually code between extremist (1) and non-extremist (0) content. Using a series of gradient boosting models and random forests, the authors content that al ML methods of detection achieve a precision rate of around 90%. The specific classifiers used in the models are not fully detailed, but all appear related to the content of the post itself.

3.4 Influence of State-Sponsored Misinformation

Does state-online propaganda actually reach its target audiences? If so, does it influence their attitudes and behavior? While these questions now play an increasing role in the research agenda on online misinformation, the answer to them remains relatively unclear and widely contested. As a result, there continue to be calls for future work that rigorously investigates whether, how, and how much the manipulation of social media by foreign actors affects public opinion and political activity (Aral and Eckles 2019).

Boyd et al's (2018, 1) analysis of 3,500 IRA ads on Facebook suggests that this method of foreign social media influence reaches a wide audience—they estimate that over 11 million Americans were exposed to these ads. At the same time, exposure to an IRA bought ad on Facebook can be fairly innocuous—viewing the ad casually while scrolling without paying much attention to its content. Offering a more fine-grained measure of the influenced population, Spangher et al (2018) find that user interaction with these IRA propaganda was extremely low—only an average of 1 in 40,000 internet users actually clicked on an IRA item (Facebook, Twitter, URL) in a given day. While this estimate of the IRA's reach is more optimistic, the Spangher et al (2018) study is limited to IRA activity in 2017 only. Looking at the broader time period of 2015–2017, Howard et al (2019, 3) find that over 30 million users actively "shared the IRA's Facebook and Instagram posts with their friends and family, liking, reacting to, and commenting on them along the way."

Dutt et al (2018) more to a more fine-grained level of analysis by seeking assess which types of IRA ads achieved the most clicks, and consequently, which strategies of foreign social media manipulation reach the broadest audience. Their findings reveal that the IRA ads most often clicked on by Facebook users tended to have low levels of positive sentiment, focused on past events, and were more specific and personalized in nature (Dutt et al 2018, 15). Given the uneven and variable audience-reach across different IRA ads on Facebook, future work on foreign influence efforts will likely benefit from a heightened attentiveness to identifying what factors make different social media efforts more-or-less effective.

Zannettou et al (2019a, 2019b) deploy Hawkes Processes to quantify the that state-sponsored trolls have on different social media platforms. Comparing between URLs posted by Russian trolls and ordinary social media users, their study (2019a) that troll accounts are generally much less influential on Reddit, Twitter, and 4Chan than other users. IRA-associated accounts were, however, significantly more influential when it came specifically to disseminating Russian state-sponsored news URLs (e.g., RT articles), which suggests operatives have a higher level of effectiveness in promoting the Russian government's own propagandistic content than they have for promoting other online content (Znnettou 2019a, 225). In a separate study, Zannettou et al (2019b, 353) use Hawkes processes to assess whether Iranian or Russia state trolls are more influential in different online communities—finding that Russian trolls were more influential across overall (notably on Twitter and Reddit), although the Iranian trolls were comparative more effective than their Russian counterparts in promoting propagandistic content on 4Chan.

Survey research on public opinion and exposure to foreign online propaganda are proving fruitful in assessing the practical influence of this form of online misinformation. Tomz and Weeks (2020) fielded a survey experiment on a sample of 3,510 US adults, with four treatment groups varying information on foreign interference on America's future 2024 election: (1) the *endorsement group*, which was told a foreign country announced a preference for one candidate; (2) a *threat group* which combined an endorsement with an implicit threat that the other candidate winning would sour relations with the US; (3) an *operation group* which was told that a foreign government used money, information, or hacking to help their favored candidate; and (4) a *stay out* control group. Compared to the *stay out group*, 46% more respondents in the *operation group* distrusted the hypothetical election results; 26% more respondents in the *operation group* reported that they would lose faith in democracy; and 9% more respondents in the *operation group* reported that they would avoid voting. These results suggest that a public revelation of a foreign election interference campaign—albeit a hypothetical one—can have significant effects on public attitudes about the quality of democracy.

Bail et al (2020) use a survey methodology that avoids hypothetical scenarios when assessing the IRA's specific impact on American public attitudes. A sample of 1,239 American Twitter users (through YouGov) agreed to share their Twitter handles with the researchers. The Bail et al surveys were conducted as a two-wave panel, and the key independent variable was whether a respondent's Twitter account interacted with IRA online content between Wave 1 and Wave 2. The results of their study demonstrate that interacting with IRA content online has no significant effect on political attitudes or levels of affective political polarization. Accordingly, the Bail et al (2020) analysis suggests that Russian trolls did not increase discord/polarization in the US; the authors note that this null effect is likely because already-polarized users are the ones selecting into interactions with IRA content online.

Fisher (2020) uses an online survey experiment to assess the influence of Russian propaganda about the Ukraine conflict on American audiences. The 895 respondents in Fisher's sample were split into four treatment conditions: (1) *Control Group*; (2) *Information Group*—who read a short news story from Russia Today (RT) about human rights violations by Ukrainian forces; (3) *Source Group*—who read that same article and were explicitly told the sources was RT; and (4) *Intentions Group*—where were additionally told that RT is "a Russian government sponsored foreign media network whose purpose is to spread a pro-Russian message to Western audiences" (Fisher 2020, 287). The experimental results show that exposure to the RT article decreases US respondents' favorable views toward Ukraine by 10%, and further, that even revealing the source (RT) and its intentions (to spread pro-Russian views) does not ameliorate this effect.

Which Target Audiences are Most Susceptible?

It is likely that online foreign influence operations are more influential over some segments of the public than others. Shu et al (2018), for instance, find that newly registered social media users (humans, not bots) are more likely to share fake news online, as are older users, women, and users with more followings that followers. These demographic and user account attributes, along with a variety of others, may potentially play some role in conditioning the effect of foreign influence operations on target audiences. Hatton and Nielsen (2016), surveying a population of undergraduate psychology students (n=280), find that demographic factors—sex (male), and having personal ties to military or law enforcement—are associated with a higher likelihood of viewing violent ISIS videos online. While viewing ISIS propaganda is not at all equatable with supporting the organizations ideology or violent tactics, exposure to these videos may influence viewers in other ways such as inducing higher levels of psychological anxiety or distress—which Natton and Nielson (2016, 170) find is significantly more likely among female viewers, younger viewers, viewers with ties to law enforcement, and conservative viewers.

Dai and Luqiu (2020) conduct an online survey experiment to assess the influence of "camouflaged propaganda"—foreign governments buying ad space in independent media outlets to publish propagandistic ads that look like normal news stories. The survey design uses real camouflaged ads posted online by the Chinese government on the Washington Post and The Telegraph. Their analysis observes that *media literacy, age, education, income, and sex* are all unrelated to respondents' abilities to correctly identify the source of the camouflaged ad as being a foreign outlet ("China Watch"). Dai and Luqiu (2020) conclude that, overall, camouflaging propaganda in independent media is an effective influence strategy—as respondents who failed to detect "China Watch" as the article source found the content significantly more persuasive when it was encountered on *Washington Post* or *The Telegraph* than did respondents who read the same article and content on *China Daily*.

Brady (2015) argues that Chinese foreign propaganda aims to promote support for the CCP among the overseas Chinese population, and further, to increase their antipathy toward anti-CCP groups. While Brady (2015) does not present empirical public opinion data on the effectiveness of China's overseas propaganda among this target audience, her analysis contends that its efforts "have been remarkably successful... especially [among] new migrants" (Brady 2015, 53). Beyond migration status and recency, other factors related to ethnic or national identity may further condition individuals susceptibility to foreign propaganda online. Benmelech and Klor (2020) examine variation in the number of ISIS foreign fighters cross nationally, and find that the size of a country's Muslim population—and not the country's local economic conditions—strongly correlate with the number of residents who become ISIS foreign recruits. Mitts (2019) further observes the degree of anti-Muslim hostility at the local (neighborhood/municipal) level is positively and significantly associated with pro-ISIS radicalization on Twitter.[9] Taken together, these studies suggest that identity based factors—and barriers toward assimilation in different countries/localities—may moderate the influence of foreign propaganda among different target audiences.

Finally, partisan affiliation—particularly in the United States—is a potentially powerful moderator of individuals' susceptibility to state-sponsored misinformation and propaganda online.

Badawy et al (2018) find that between September–November 2016, American conservatives retweeted Russian trolls with a significantly higher frequency than American liberals.[10] Accordingly, Badawy et al (2018, 258) conclude that while a ideologically diverse population of Twitter users were exposed to Russian online trolls, "it was mainly conservatives who helped amplify their message." Using the same data on election related tweets between September–November 2016, Badawi et al (2019) show that political ideology (conservativism) is the most predictive factor explaining which users do—and do not—retweet Russian trolls. Gallacher et al (2018, 6) find that US military personnel (active-duty and non-active) exhibit "significant and persistent interactions... [with] a broad network of Russia focused accounts... [and] these

[9] Mitts(2019) bases these findings on an original data set tracking ISIS activist and follower accounts on Twitter (web scraped using real-time lists provided by anti-ISIS hacking groups). Mitts then uses a spatial label propagation algorithm to predict Twitter users' locations based off of information about their online social network. Finally, local level of anti-Muslim hostility is proxied using an indicator for the local vote share of far-right parties.

[10] This finding is based upon authors' analysis of a data set coding 43 million elections-related posts on Twitter by 5.7 million different users. Individual ideology is classified based upon the news sources that different users share online.

interactions are often mediated by proTrump users."[11] The findings of these studies offer support for the notion that partisanship and polarization—particularly among conservative Americans—is associated with higher levels of influence by foreign propaganda online.

At the same time, the degree to which partisanship conditions susceptibility to foreign propaganda may not be constant across different policy issues or topics. Carter and Carter (2021) conduct a survey experiment assessing the influence of RT propaganda in the American public. They find that exposure to RT propaganda online has no influence on American's faith in democratic institutions (e.g., presidential approval, trust in government)—a conclusion that is consistent with the notion that Americans' domestic policy views begin "more calcified than those on foreign policy" (Carter and Carter 2021, 70). By contrast, exposure to RT propaganda was associated with significant differences in foreign policy views—with the treatment group (propaganda exposure) believing that the American public should reduce is presence and activity on the world stage. This effect was consistent across partisan lines, and it persisted even when respondents were explicitly told that RT is financed by the Russian government (Carter and Carter 2021, 49). Consequently, partisan identity and polarization may moderate the influence of foreign propaganda less for some policy issues than others.

[11] Gallacher et al (2018) construct an original data set of 28,467 Twitter accounts that follow military junk-news sources, and track the activity of those accounts between April–May 2017.

References

[1] Aly, Anne, Stuart Macdonald, Lee Jarvis, and Thomas M. Chen. (2017). "Introduction to the Special Issue: Terrorist Online Propaganda and Radicalization." *Studies in Conflict & Terrorism* 40(1): 1–9.

[2] Aral, Sinan, and Dean Eckles. (2019). "Protecting Elections from Social Media Manipulation." *Science* 365(6456): 858–861.

[3] Badawy, Adam, Emilio Ferrara, and Kristina Lerman. (2018). "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign." *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 258–265.

[4] Badawy, Adam, Kristina Lerman, and Emilio Ferrara. (2019). "Who Falls for Online Political Manipulation?" *Companion Proceedings of The 2019 World Wide Web Conference* 162–168.

[5] Bail, Christopher A., Brian Guay, Emily Maloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. (2020). "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017." *Proceedings of the National Academy of Sciences* 117(1): 243–250.

[6] Barron-Cedeno, Alberto, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. (2019). "Proppy: A System to Unmask Propaganda in Online News." *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1): 9847–9848.

[7] Benmelech, Efraim, and Esteban F. Klor. (2020). "What Explains the Flow of Foreign Fighters to ISIS?" *Terrorism and Political Violence* 32(7): 1458–1481.

[8] Boyd, Ryan L., Alexander Spangher, Adam Fourney, Besmira Nushi, Gireeja Ranade, James Pennebaker, and Eric Horvitz. (2018). "Characterizing the Internet Research Agency's Social Media Operations During the 2016 US Presidential Election Using Linguistic Analyses." *ArXiv Preprint* 1–9.

[9] Bradshaw, Samantha and Philip Howard. (2017). "Troops, Trolls, and Troublemakers: A Global Inventory of Organized Social Media Manipulation." *Technical Report, Oxford Internet Institute.*

[10] Brady, Anne-Marie. (2015). "Authoritarianism Goes Global (II): China's Foreign Propaganda Machine." *Journal of Democracy* 26(4): 51–59.

[11] Broniatowski, David A., Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. (2018). "Weaponized Health Communication:

Twitter Bots and Russian Trolls Amplify the Vaccine Debate.” *American Journal of Public Health* 108(10): 1378–1384.

[12] Carter, Erin Baggott, and Brett L. Carter. (2021). “Questioning More: RT, Outward-Facing Propaganda, and the Post-West World Order.” *Security Studies* 30(1): 49–78.

[13] Dai, Yaoyao, and Luwei Luqiu. (2020). “Camouflaged Propaganda: A Survey Experiment on Political Native Advertising.” *Research & Politics* 7(3): 1–10.

[14] Dutt, Ritam, Ashok Deb, and Emilio Ferrara. (2018). “ “Senator, We Sell Ads”: Analysis of the 2016 Russian Facebook Ads Campaign.” *International Conference on Intelligent Information Technologies* 151–168.

[15] Etudo, Ugo, Victoria Y. Yoon, and Niam Yaraghi. (2019). “From Facebook to the Streets: Russian Troll Ads and Black Lives Matter Protests.” *Proceedings of the 52nd Hawaii International Conference on System Sciences* 894–901.

[16] Farkas, Johan, and Marco Bastos. (2018). “IRA Propaganda on Twitter: Stoking Antagonism and Tweeting Local News.” *Proceedings of the 9th International Conference on Social Media and Society*: 281–285.

[17] Farwell, James P. (2014). “The Media Strategy of ISIS.” *Survival* 56(6): 49–55.

[18] Fernandez, Alberto M. (2015). “Here to Stay and Growing: Combating ISIS Propaganda Networks.” *Center for Middle East Policy at Brookings*.

[19] Fisher, Aleksandr. (2020). “Demonizing the Enemy: The Influence of Russian StateSponsored Media on American Audiences.” *Post-Soviet Affairs* 36(4): 281–296.

[20] Gallacher, John D., Vlad Barash, Philip N. Howard, and John Kelly. (2018). “Junk News on Military Affairs and National Security: Social Media Disinformation Campaigns Against US Military Personnel and Veterans.” *arXiv preprint* 1–6.

[21] Gertsel, Dylan. (2016). “ISIS and Innovative Propaganda: Confronting Extremism in the Digital Age.” *Swathmore International Relations Journal* 1(1): 1–9.

[22] Gorwa, Robert. (2017). “Computational Propaganda in Poland: False Amplifiers and the Digital Public Sphere.” *Working Paper, Oxford Project on Computational Propaganda.*

[23] Hanlon, Bradley. (2018). “It’s Not Just Facebook: Countering Russia’s Social Media Offensive.” *German Marshall Fund of the United States.*

[24] Hansen, Isabella and Darren J. Lim. (2019). “Doxing Democracy: Influencing Elections via Cyber Voter Interference.” *Contemporary Politics* 25(2), 150–171.

[25] Hatton, Arthur T., and Michael E. Nielsen. (2016). “ ‘War on Terror’ in our Backyard: Effects of Framing and Violent ISIS Propaganda on Anti-Muslim Prejudice.” *Behavioral Sciences of Terrorism and Political Aggression* 8(3), 163–176.

[26] Hegelich, Simon, and Dietmar Janetzko. (2016). “Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet.” *Tenth International AAAI Conference on Web and Social Media* 579–582.

[27] Helmus, Todd C., Elizabeth Bodine-Baron, Andrew Radin, Madeline Magnuson, Joshua Mendelsohn, William Marcellino, Andriy Bega, Zev Winkelman. (2018). “Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe.” *Rand Corporation.*

[28] Howard, Philip N., Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille Franc¸ois. (2019). “The IRA, Social Media and Political Polarization in the United States, 2012–2018.” *Oxford University, Computational Propaganda Research Project.*

[29] Im, Jane, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. (2020). “Still Out There: Modeling and Identifying Russian Troll Accounts on Twitter.” *12th ACM Conference on Web Science* 1–10.

[30] Kim, Young Mie, Jordan Hsu, David Neiman, Colin Kou, Levi Bankston, Soo Yun Kim, Richard Heinrich, Robyn Baragwanath, and Garvesh Raskutti. (2018). “The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on Facebook.” *Political Communication* 35(4), 515–541.

[31] Klausen, Jytte, Eliane Tschaen Barbieri, Aaron Reichlin-Melnick and Aaron Y. Zelin. (2012). “The YouTube Jihadists: A Social Network Analysis of Al-Muharijoun’s Propaganda Campaign.” *Perspectives on Terrorism* 6(1), 36–53.

[32] Klausen, Jytte. (2014). “Tweeting the Jihad: Social Media Networks of Western Fighters in Syria and Iraq.” *Studies in Conflict & Terrorism* 38(1), 1–22.

[33] Linvill, Darren L., Brandon C. Boatwright, Will J. Grant, and Patrick L. Owen. (2019). “ “THE RUSSIANS ARE HACKING MY BRAIN!” Investigating Russia’s Internet Research Agency Twitter Tactics During the 2016 US Presidential Campaign.” *Computers in Human Behavior* 99, 292–300.

[34] Martin, Diego A., and Jacob N. (Working Paper). “Trends in Online Foreign Influence Efforts.”

[35] McFaul, Michael and Bronte Kass. (2019). “Understanding Putin’s Intentions and Actions in the 2016 US Presidential Election,” in *Securing American Elections: Perceptions for*

Enhancing the Integrity and Independence of the 2020 U.S. Presidential Election and Beyond, ed. Michael McFaul. Stanford University Cyber Policy Institute.

[36] Mitts, Tamar. (2019). “From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West.” *American Political Science Review* 113(1): 173–194.

[37] Mussiraliyeva, Shynar, Milana Bolatbek, Batyrkhan Omarov, Zhanar Medetbek, Gulshat Baispay, and Ruslan Ospanov. (2020). “On Detecting Online Radicalization and Extremism Using Natural Language Processing.” *2020 21st International Arab Conference on Information Technology (ACIT)* 1–5.

[38] Orrtung, Robert W. and Elizabeth Nelson. (2019). “Russia Today’s Strategy and Effectiveness on YouTube.” *Post-Soviet Affairs* 35(2), 77–92.

[39] Pomerantsev, Peter. (2015). “Authoritarian Goes Global (II): The Kremlin’s Information War.” *Journal of Democracy* 26(4), 40–50.

[40] Romm, Tony and Kurt Wagner. (6 October, 2017). “Silicon Valley’s Russian Ads Problem, Explained.” *Vox News*: https://www.vox.com/2017/10/6/16419388/ facebook-google-twitter-russia-ads-2016-election-president-donald-trump.

[41] Rudner, Michael. (2017). “ “Electronic Jihad”: The Internet as Al Qaeda’s Catalyst for Global Terror.” *Studies in Conflict & Terrorism* 40(1): 10–23.

[42] Shu, Kai, Suhang Wang, and Huan Liu. (2018). “Understanding User Profiles on Social Media for Fake News Detection.” *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* 430–435.

[43] Stamos, Alex. (6 September, 2017). “An Update on Information Operations on Facebook.” *Facebook Newsroom*: https://about.fb.com/news/2017/09/ information-operations-update/.

[44] Spangher, Alexander, Gireeja Ranade, Besmira Nushi, Adam Fourney, and Eric Horvitz. (2018). “Analysis of Strategy and Spread of Russia-sponsored Content in the US in 2017.” *arXiv Preprint*.

[45] Stewart, Leo G., Ahmer Arif, and Kate Starbird. (2018). “Examining Trolls and Polarization with a Retweet Network.” *Proceedings of AcM WSDM Workshop on Misinformation and Misbehavior Mining on the Web* 70: 1–6.

[46] Tomz, Michael, and Jessica LP Weeks. (2020). “Public Opinion and Foreign Electoral Intervention.” *American Political Science Review* 114(3): 856–873.

[47] Tundis, Andrea, Ahmed Ali Shams, and Max Muhlh¨ auser. (2020). “Concepts of a Pyramidal¨ Model for Assessing Internet-Based Terrorist Propaganda.” *2020 IEEE 19th International Symposium on Network Computing and Applications* 1–4.

[48] Walter, Dror, Yotam Ophir, and Kathleen Hall Jamieson. (2020). “Russian Twitter Accounts and the Partisan Polarization of Vaccine Discourse, 2015–2017.” *American Journal of Public Health* 110(5): 718–724.

[49] Zannettou, Savvas, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. (2019a). “Disinformation Warfare: Understanding StateSponsored Trolls on Twitter and Their Influence on the Web.” *Companion Proceedings of the 2019 World Wide Web Conference* 218–226.

[50] Zannettou, Savvas, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. (2019b). “Who Let the Trolls Out? Towards Understanding StateSponsored Trolls.” *Proceedings of the 10th ACM Conference on Web Science* 353–362.

4. Vulnerability to Misinformation in the U.S.

Prepared primarily by SS

This section addresses research on the following question: *How did the American public become so vulnerable to online misinformation and propaganda*? Section 1 covers scholarship on the most prominent answer to this question—that partisan polarization amplifies receptivity to misinformation. Section 2 then highlights research on psychology that indicates emotional reasoning has an additional—and sometimes independent—effect on misinformation vulnerability. Finally, Section 3 scrutinizes how the structure and attributes of online social networks may facilitate a heightened susceptibility to online misinformation among the U.S. public.

4.1 Partisan Polarization and Misinformation Vulnerability

In comparative perspective, the online misinformation environment in the United States is somewhat distinctive in the degree to which it is colored by partisanship. Humprecht (2019) contrasts the content of online "fake news" stories in four Western democracies (US, UK, Germany, and Austria),[1] Her content analysis identifies a divergence in the nature of online misinformation in the English-speaking (US/UK) versus German-speaking (Germany/Austria) countries: in the former, the bulk of misinformation content is patently partisan and directed against political actors; in the latter, most misinformation comes in the form of sensationalist stories and is commonly targeted at immigrants. While online political misinformation is shared across all four cases, the US and UK were unique in the extent that such information was overtly colored by partisanship.

[1] This comparison rests on an analysis of up to 100 false news stories published by two fact-checking websites in each country between June 2016 and September 2017 (n=651 total stories).

The partisan nature of online misinformation in the United States is unlikely to be a coincidence, but instead a byproduct of a broader relationship between political polarization and misinformation vulnerability. Partisan identity and attitudes in the US map onto belief in online misinformation as well as further propagation of such misinformation, which is likely what makes partisan-charged misinformation items more pervasive in the American context. Osmundsen et al (2020), for instance, argue that "partisan polarization is the primary psychological motivation behind 'fake news' sharing on Twitter."[2] The authors recruited a sample of 8,741 respondents to take a survey measuring their *partisanship*, *political cynicism*, *in-party love*, and *out-party hate* alongside a variety of other factors.[3] After completing the survey, 2,337 respondents agreed to share their public Twitter handles with the research team. Analyzing the survey results in conjunction with respondents' public positing history on Twitter, Osmundsen et al (2020) find that *political cynicism, conservative partisanship*, and level of *out-party hatred* were all associated with a significantly higher frequency of fake news sharing online. Moreover, respondents who reported high-levels of hate toward the out-party disproportionately shared online information (true and untrue) that was specifically useful in derogating members of that party (Osmundsen et al 2020, 1012). These results suggest that susceptibility to misinformation in the U.S.—or at least willingness to propagate it—is related to both partisan identity (conservative) and polarization (out-party hate).

Nikolov et al (2020) similarly argue that partisan polarization drives susceptibility to online misinformation, but they go farther in highlighting how it is not just right-leaning social media users who display a heightened propensity to share misinformation online. After scraping a 10% random sample of public posts containing URLs from Twitter's API, the authors then restrict their analysis to users who: (1) actively consume online political news;[4] and (2) were vulnerable to

[2] This is their article's title verbatim.

[3] Survey conducted between December 2018 and January 2019.

[4] This was operationalized as users who shared at least 10 links from news sources with a known political valence in the study period (June 1–June 30, 2017).

misinformation.[5] The resulting data set comprised a sample of 1,398,552 Tweets, and individual users' partisanship was measured by tracking post with links to 500 different news organizations with different political leanings (Fox, CNN, etc.). While right-leaning partisans were indeed more likely to share misinformation in public Tweets, a similar (though weaker) trend was displayed among left-leaning users as well. Overall, both partisan groups (right/left) were significantly more likely to spread online misinformation than were ideological moderates (Nikolov et al 2020, 5). Thus, the most ideologically polarized social media users appeared to also be the most vulnerable to online misinformation.

Ribeiro et al (2017) argue that political polarization in the United States even colors what individuals perceive as "fake news" in the first place. Their analysis is based upon two data sets compiled by scraping the Twitter API between May 2017–August 2017: (1) a data set on general political views used to compute user ideology based off of tweets referencing US politics; and (2) a data set collecting tweets including hashtags identifying some content as "fake news"—#FakeNews, #AlternativeFacts, etc. Their results indicate that the degree of political polarization among Twitter users is significantly higher in online conversations about what does—and does not—constitute "fake news" than it is in other conversations about politics in which "fake news" is not an explicit allegation.[6] In this sense, partisan polarization can facilitate susceptibility to misinformation online by producing correspondingly polarized views on what forms of content are even appropriately classified as "misinformation"—with views on this themselves diverging along partisan lines.

[5] This was operationalized as users who shared at least one link from a low-quality news source, and resulted in excluding 5% of right-leaning users and 30% of left-leaning users.

[6] Online discussions about politics were also highly polarized, but the degree of polarization was still significantly lower in political tweets that did not reference "fake news" as an explicit theme.

Polarization, Media/Political Elites, and Misinformation

Elite messaging interacts with partisan polarization to produce a heightened susceptibility to misinformation among the American public. Nyhan (2010) analyzes misinformation in the content of public debates over health care reform under Bill Clinton (myth = loss of doctor choice) and Barack Obama (myth = death panels). In each case, Nyhan (2010) concludes that elite polarization contributed to misinformation by reinforcing a partisan divide in factual beliefs. Specifically, conservative media and political elites endorsing these various myths (e.g., Sean Hannity, Rush Limbaugh, Sarah Palin, Glenn Beck, Chuck Grassley) fed into factually misinformed beliefs across that American public that mapped onto partisan lines. Analyzing a 2009 Pew survey on news interest, Meirick (2013) similarly finds that Republicans who tuned into Fox News or right-wing radio stations were significantly more likely to hold misinformed beliefs about death panels because of the politically biased messaging provided through many of these outlets.

Moreover, just as highly-partisan individuals are more likely to spread misinformation (Osmundsen et al 2020; Nikolov et al 2020), highly-partisan media organizations are themselves more likely to produce it. Shrestha et al (2019) examine how the partisan-bias of 1,785 different publishers relates to the credibility of news content that they disseminate.[7] Highly-partisan media organizations were observed to be more likely to write fake news, with 80% of publishers in the "right bias" group generating content with an overall "mixed credibility" level, and also 40% of publishers in the "left bias" group having a mixed credibility level. By comparison no more than 12.5% of left-center, right-center, and moderate publishers were found to have such a mixed-credibility in their news content—reflecting a greater tendency of these outlets to avoid biased/unreliable sources and use proper sourcing methods in their articles.

At the same time, the growing misinformation problem in the U.S. is not just a result of individuals trusting non-credible partisan news. It also reflects a partisan divide in the extent to

[7] Publisher partisanship operationalized using MediaBias webpage and credibility operationalized using FactCheck webpage.

which individuals actually distrust reliable news information. Benkler et al (2018, 20–21) argue that the explosion of misinformation in the U.S. (i.e., the "post-truth" age) has a distinctly partisan shape because right-wing audiences are uniquely distrusting of conventional media sources, hostile toward professional journalism, and correspondingly susceptible to misinformation in the form of conspiracies, click-bait, and Russian propaganda.

Uscinski et al (2016) combine an observational survey analysis of 1,230 Americans in 2012 with a follow-up survey experiment on media bias and political attitudes (n=1,015). Their observational results shows that Democrats and Republicans are about equally disposed to conspiratorial thinking in general (about half of each group believes in various measures of "secret elite control" of the country behind closed doors). Yet, their experimental results—being more specifically focused on media distrust—show a clear partisan divide in susceptibility to conspiracies: Republicans are 44% more likely than Democrats to believe that the mainstream media engages in biased election reporting due to a grand political conspiracy. Hence, Republican distrust of mainstream media risks feeding into a conspiratorial world view in which the provision of information is itself viewed in inherently partisan terms.

In addition to distrust of media sources that provide access to reliable information, Republicans in the United States—particularly throughout the Trump presidency—were routinely exposed to discourse by co-partisan elites that created and amplified political misinformation. Hornsey et al (2020), for instance, explore whether anti-vaccine misinformation spread by Donald Trump affected public views of vaccine efficacy, and found that Trump voters were particularly receptive to this form of misinformation from the President. The authors use an MTurk survey experiment (n=518) fielded in two waves one-week apart: in each wave, respondents reported their views on the safety and efficacy of vaccines—and between the two waves, half of the respondents were exposed to real tweets by Donald Trump that contested the efficacy of vaccinations. Trump supporters (but not other respondents) exposed to these anti-vaccine tweets grew significantly more concerned about vaccines between waves 1 and 2, indicating that elite misinformation messaging by the President translates into disparate acceptance of that misinformation along partisan lines.

Beyond public receptivity to misinformation, partisan polarization is also associated with lower audience costs imposed upon elites who spread political misinformation through their rhetoric. Swire-Thompson et al (2020) conduct a survey experiment through MTurk in which two different political candidates (Bernie Sanders and Donald Trump) made verifiably false claims[8] which subsequently confirmed of debunked for respondents through a fact-check. Their results show that: (1) supporters of each candidate initially believe the misinformation to be true; (2) fact-checking misinformation reduces belief in the accuracy of that information among respondents;[9] but (3) factchecks flagging clear misinformation did not have any effect on respondents' overall views toward either political candidate. In an earlier survey experiment, Swire et al (2017) similarly found that: (1) Republican supporters of Trump believed misinformation items at significantly higher rates if they were explicitly attributed to Trump as opposed to non-attributed; (2) when misinformation by Trump was corrected, Republicans reduced their belief in the information but did not change their voting preferences or punish Trump for spreading falsehoods; and (3) after a one-week delay, respondents re-believed misinformation after it was previously corrected—reverting back to their original partisan assumptions.[10]

Further supporting these conclusions, Nyhan et al (2020) use a survey experiment introducing journalistic fact-checks of misinformation claims made by Trump during the 2016 Republican Convention and presidential debates. Their findings demonstrate: (1) that fact-checking

[8] In the control condition, each candidate made an equal number of true and false claims. In the treatment condition, each candidate made disproportionately more false claims than true ones.

[9] Trump supporters reduced their belief in misinformation after it was fact-checked, but they did so to a significantly lower degree than did non-supporters. Sanders supporters and non-supporters, by contrast, reduced belief in misinformation at similar rates after it was fact-checked.

[10] This earlier experiment only included Donald Trump, and it was conducted in 2015 prior to the Iowa caucus. The design had respondents report their support for Trump and then level of belief in misinformation statements made by Trump (attributed or not), then those misinformation items were factually corrected, and respondents re-rated their belief levels and support for Trump (either right away, or after a one-week delay).

misinformation by Trump improved the accuracy of respondents factual beliefs, even among his supporters; but (2) these fact-checks had no measurable effect on attitudes toward Trump. Taken together, the results of these studies suggest that misinformation can be effective (particularly when not factchecked), and further, that political elites face little costs for spreading misinformation because views toward them are fairly entrenched and not their public support is insensitive to getting caught in a clear lie.

Political polarization prevents also filters fact-checks of misinformation by political candidates, such that many fact-checking messages will fail to reach individuals exposed to false claims and correct their views. Shin and Thorson (2017) investigate how Twitter users shared fact-checks that the Tampa Bay Times, Washington Post, and Factcheck.org made on remarks from Barack Obama and Mitt Romney during the 2012 presidential debates. Their results show that "partisans selectively share fact-checking messages that cheerlead their own candidate and denigrate the opposing party's candidate, resulting in an ideologically narrow flow of fact checks to their followers" (Shin and Thorson 2017, 233). Thus, another reason political elites get away with spreading misinformation in today's polarized environment is that partisan blinders mean that fact-checks do not spread to—and are not shared by—their existing supporters.

Partisan Identity and Misinformation

Although partisan polarization facilitates the spread of misinformation throughout right- and leftleaning segments of the U.S. public, there are good reasons to doubt that it does so equally in each camp. Indeed, a number of recent studies suggest that American conservatives (e.g., Republicans, Trump-supporters) are significantly more susceptible to believing and spreading online misinformation.

Guess et al (2018) surveyed 2,525 Americans and received consent to collected web browsing data from their computers between October 7–November 14, 2016. In assessing the relationship between partisanship and fake news consumption, they find that while 25% of respondents visited at least one fake news website in the study period, most visits were logged by US conservatives. Trump supporters in particular "visited the most fake news websites, which were overwhelming pro-Trump. However, fake news consumption was heavily concentrated among a small group— almost 6 in 10 visits to fake news websites came from people with the most conservative online information diets" (Guess et al 2018, 1). Havey (2020, 319) analyzes a sample of 4,101 tweets related to COVID-19 misinformation (e.g., hydroxychloroquine, bleach, Bill Gates) and concludes that across the board, "conservatives are more likely than their liberal peers to believe in and push conspiracy theories" about COVID-19.[11]

To analyze how partisan identity relates to election misinformation, Grinberg et al (2019) match 16,442 Twitter users with U.S. voter registration records and analyze their account activity between August and December 2016. Their study identifies a similar partisan gap in misinformation receptivity: "fewer than 5% of people on the left or in the center ever shared any fake news [political] content, yet 11 and 21% of people on the right and extreme right did, respectively" (Grinberg et al 2019, 3).[12]

Badawy et al (2018) analyze a sample of 43 million election-related tweets between September– November 2016. Their sample identifies tweets by Russian trolls using data released by the US House Intelligence Committee, and for non-troll users ideology is inferred using the news sources shared on their feed. American conservatives in their sample retweeted misinformation by Russian trolls significantly more often by liberals, which leads Badawy et al (2018, 258) to conclude that "although an ideologically broad swath of Twitter users were exposed to Russian trolls... it was mainly conservatives who helped amplify their message." In a follow-up analysis using this same data set, Badawy et al (2019) employ machine learning models to identify the set of factors most predictive of Twitter users who do/don't share Russian troll misinformation online. The results of this study indicate that "political ideology is the most

[11] Twitter user ideology was computed using Barbera's Tweetscores R package, which estimate a Twitter user's´ political ideology based on their association with elite users (e.g., Obama, Clinton, Trump, Sanders, Hannity, etc.).

[12] These percentages are restricted to the sample of Twitter users who shared any political content during the election; those who did not share any political content are excluded. User ideology was computed by estimating the similarity of each user's Twitter news feed to those of registered Democrats or Republicans.

predictive feature," with conservative users sharing Russian troll posts at a much higher level of frequency (Badawy et al 2019, 166).

In the context of the MacronLeaks online misinformation campaign leading up to the 2017 French election, Ferrara (2017) shows—perhaps surprisingly—that right-wing Americans were the most active social media users in boosting misinformation on this topic. Analyzing 17 million Twitter posts between April 27–May 7, 2017, Ferrara (2017, 1) finds that foreigners—rather than French users—with a preexisting interest in alt-right topics (e.g., #MAGA, #Trump) were most actively engaged in spreading MacronLeaks information online. This provides evidence that rightleaning Americans have a higher propensity to consume and share online misinformation, even about topics abroad that do not have as clear of a direct relevance to U.S. domestic politics.

Studies that find American conservatives to be more active in consuming and spreading online misinformation introduce a degree of conditionality into the argument that partisan polarization fuels modern misinformation vulnerabilities. There is evidence for this dynamic on both the left and the right (Shrestha et al 2019; Nikolov et al 2020), but the magnitude of the effect appears somewhat greater among right-leaning segments of the American public in practice.

Partisan Polarization and Misinformation: Mechanisms

How, exactly, does partisan polarization make American audiences more vulnerable to misinformation? Motivated reasoning and belief consistency are often identified as a key mechanism producing this relationship.

Schaffner and Roche (2016) examine how motivated reasoning conditions factual beliefs about politics using a 2010–2012 Cooperative Congressional Election Survey of 11,000 U.S. respondents. Their analysis of this survey data is restricted to a two-week period immediately before and after the announcement of a significant decline in unemployment a month before the 2012 presidential election (DV = respondents estimates of the real unemployment rate). Their results show that after the US jobs report was announced in 2012, Democrats uniformly updated their assessment of the unemployment rate in the direction of greater accuracy. By contrast, while some Republicans became more accurate in their estimates, an even larger share of Republicans updated their estimates of unemployment to be even higher and less accurate after the jobs report was announced.[13] Schaffner and Roche (2016) interpret these results as indicating that partisan engagement in motivated reasoning explains differential reactions to the job report announcement—with Democrats updating toward accuracy in order to cheerlead Obama, and Republicans overestimating unemployment in order to avoid crediting Obama with an economic success. In this sense, Republicans likely failed to assimilate accurate information precisely because it conflicted with their own political priors.

Kuklinski et al (2020) also find partisan-motivated reasoning to be at play in explaining why Americans are misinformed about welfare spending in the United States. Analyzing respondent estimates of welfare spending in a telephone survey, the authors conclude that strong partisans are simultaneously the must misinformed and most confident in their beliefs about how much of the US budget gets devoted to social welfare programs (Kuklinski et al 2020, 799). Americans, thus, are not simply uniformed about social policy, but confidently misinformed in ways that broadly map onto partisan priors. Using a small-scale experiment on undergraduate students (n=83), Moravec et al (2019) find that flagging fake news has no effect on respondents judgments regarding the accuracy of information. Instead, the authors conclude that motivated reasoning lead respondents to believe news headlines that align with the political priors even when those headlines are flagged as being misinformation.

Using a method of “challenge interviews” counter-arguing against misinformation (n=84), Prasad et al (2009) find that preserving belief-consistency drove US Republicans to resist the correction of misinformation alleging Saddam Hussein was directly involved in the 9/11 terrorist attacks. Republicans who previously reported believing Saddam Hussein was involved deployed

[13] Republicans took more time to answer this question, suggesting they did more research and spent more cognitive resources in formulating their answers. And Republicans who took the most time to answer the question provided, on average, the least accurate responses.

a variety of tactics to resist conflicting information: ignoring, counter-arguing, inferred justification (“some evidence must exist, or else we wouldn’t have invaded Iraq”), or disputing the need for opinions to be grounded in facts altogether.

Bullock et al (2013) challenge the notion that motivated reasoning drives a partisan split over belief in misinformation. In a 2008 CCES survey of 626 Americans, they ask a series of factual questions about politics and provide half of the respondents (randomly assigned) a direct financial incentive to respond accurately (each correct question gets one entry into a drawing for a $200 gift card). Their results show that being offered a small financial incentive decreased the partisan gap by half—which the authors interpret as providing evidence for a “cheerleading” mechanism whereby strong-partisans falsely profess their belief in polarizing misinformation items.

Peterson and Iyengar (2021) reopen this question and aim to differentiate between whether partisan disagreements over factual information are either: (1) genuine differences of opinion; or (2) insincere cheerleading—claiming to believe something one actually does not because it benefits one’s preferred party if true. They ask respondents a series of politically salient factual questions (do most scientists believe in climate change, did Michael Cohen plead guilty to campaign crimes, did Obama wiretap Trump, etc.) and offered a treatment-group a 50cent incentive to provide correct answers. On each topic, respondents were allowed to read articles from left sources (MSNBC, Huff Post), right sources (Fox, Brietbart), mainstream sources (CNN, NYT), or expert sources (e.g., Bureau of Labor Statistics on unemployment rate question). They found that offering financial incentives had no impact on partisan information searching, and only very minimally reduced partisan divides in professed factual beliefs (a 60%–70% partisan divide over factual information still persisted). Overall, their findings “support the motivated reasoning interpretation of misinformation; partisans seek out information with a congenial slant and sincerely adopt inaccurate beliefs to cast their party in a favorable light” (Peterson and Iyengar 2021, 133).

4.2 Emotionality and Misinformation Vulnerability

While much research on the surge of political misinformation in the United States identifies partisan polarization as a key driving force, recent work in political psychology positions emotional states as a potential moderator in the relationship between ideology and misinformation susceptibility. Emotions like fear, anger, stress, and anxiety in individuals may condition the degree to which they process information in a highly-partisan or open-minded manner. If such emotions are indeed moderators of misinformation acceptance, then heightened levels of hate/anger/anxiety in the U.S. public may feed into the modern misinformation problem—while getting the US public to just chill out a bit may do work in alleviating that problem.

Weeks (2015) conducts an online experiment in the U.S. (n=768) that randomly assigned respondents to first write something about immigration/the death penalty that made them: (1) angry and anxious; or (2) relaxed. Respondents then read a news article discussing how there were misperceptions circulating about the topic they chose. That article included a series of inaccurate claims, and varied whether those claims were: (1) corrected or not-corrected; and (2) attributed to Congressional Democrats or Republicans. Weeks (2015, 699) finds that "anger encourages partisan, motivated evaluation of uncorrected misinformation that results in beliefs consistent with the [respondent's] supported political party... [while] exposure to corrections improves belief accuracy, regardless of emotion or partisanship." In this sense, angry partisans—and not just strongpartisans—may be uniquely more vulnerable to uncorrected misinformation.

Analyzing the relationship between partisanship, emotion, and misinformation, Freiling et al (2021) similarly conduct an experiment that first primes participants to write about something that either makes them *extremely anxious* or *extremely relaxed*. Subsequently, all respondents were provided six news posts on Facebook about COVID-19 (shared either by Fox or MSNBC depending on treatment condition). Two posts contained accurate information, two posts contained misinformation, and two posts contained fact-checks. Their results show that induced *anxiety* made respondents significantly more likely to share misinformation claims, and further, that the effect of anxiety was notably higher in Republicans. While all Republicans were more likely to believe and share COVID-19 misinformation, it was specifically the high-anxiety Republicans who were most believing and most willing to spread this false content online.

In a small experiment on study on university students (n=97) that induced anger in a treatment condition by using a disorganized, insulting, and dismissive experimenter, Greenstein and Franklin (2020) find that angry participants were more susceptible to believing misinformation about a movie clip they were shown. On a more clearly political topic, Valentino et al (2008) examine how inducing different emotional states in a lab experiment (n=408, conducted at University of Michigan) affected their willingness to seek out new information about political campaigns. The authors found that participants primed to think of topics which made them angry spent significantly less time seeking out information about campaigns online—compared to both the control condition and other emotional treatments (anxiety, enthusiasm).

Moving beyond analyzing individual emotions separately, Martel et al (2020) conduct two separate surveys that assess how respondents' emotional reasoning affect their susceptibility to online misinformation. In the first study, respondents were first asked a battery of questions about how strongly they felt X/Y/Z at the moment they were taking the survey. In the second study, respondents were randomly primed to consider either the virtues of *emotional induction* or *reasonbased induction*. In the first study, highly emotional respondents—regardless of emotion type or valence—were significantly more believing in fake news headlines presented to them. And in the second study, respondents in the *emotional induction* condition exhibited a significantly higher level of belief in the fake news headlines.

While external validity is a concern with all of these experiments, particularly those on university samples, they reflect a body of work in political psychology that associates misinformation vulnerabilities not just to political ideology—but also to discrete emotional states. While it seems quite difficult to assume the American public is broadly becoming more emotional, there may be a justification for a more limited form of this assumption during the

COVID-19 period—in which anxiety, stress, and perhaps even anger were heightened due to a whole series of different stresses that affected the population writ large.

4.3 Social Network Features and Misinformation Vulnerability

Technological change in patterns of news consumption over the past few decades may further be contributing to America's growing vulnerability to political misinformation. Indeed, acceptance of misinformation by the US public may be exacerbated by heightened reliance on social media, segmentation of users into homogeneous echo chambers, filter bubbles that prevent ideologically incongruent misinformation from penetrating those echo chambers. Flaxman et al (2016) examine the web-browsing histories for 50,000 American adults who actively read the news (derived from a sample of 1.2 million US users) between March and May 2013. They find that Americans who get their news through social media or web-searches are significantly more polarized than those who directly visit news sites.[14] Moreover, individuals broadly read news publications that were ideologically similar to one another and encounter articles from only one side of the partisan spectrum, leading Flaxman et al (2016, 317) to conclude that "many—indeed nearly all—users exist in so-called echo chambers."

Allcott and Gentzkow (2017) analyze the share of traffic for different online news outlets (690 top news sites in the USA, and 65 fake news sites), finding that only 10.1% of traffic to top-news sites comes through social media while 41.8% of traffic to fake news sites comes through social media. Allcott and Gentzkow (2017, 221) identify three reasons why increased reliance on social media is magnifies the extent the misinformation problem in the United States: (1) entering the social media news market has small costs, and click-bait tactics have high short-term payoffs— which attract and incentivize non-credible publishers; (2) the format of social media—

[14] Polarization is measured by the political slant of news outlets users interact with. Political slant of news outlets measured by the fraction of the outlet's readership who voted for the Republican party in 2012—tracking the location of each webpage view by IP address to the county-level, and proxying using the county's political composition.

viewed on phones—makes it harder for users to judge an article's veracity; and (3) social media networks are ideologically segregated, so people receive ideologically congruent content but not necessarily true content.

Tornberg (2018) investigates whether the presence of echo chambers in social media networks are a causal factor driving the increased virality of online misinformation. Tornberg devises a theoretical as opposed to empirical model of how misinformation spread throughout a social network— modeling it as a complex diffusion in which each node has an activation threshold randomly assigned that reflects the fraction of misinformation nodes it must be connected to in order to activate itself. Introducing an echo chamber into this diffusion model involves adding a portion of densely connected nodes that with a high level of opinion homophily—their activation thresholds are set lower to reflect greater opinion consistency. Running a series of simulations on the diffusion of misinformation in networks with an echo chamber, Tornberg (2018, 16) concludes that a global misinformation cascade throughout the broader network is more likely in the presence of an echo chamber through a band-wagoning effect: "when misinformation resonates with the views of an echo chamber, the chamber can function as an initial platform from which the diffusion can occur globally through weak ties." Thus, to the extent that heightened reliance on social media in the U.S. produces information sharing networks with more echo chambers, the capacity for misinformation to cascade online is amplified.

Anspach and Carlson (2020) argue that information gathering shortcuts commonly used by individuals who get news through social media further facilitate the spread of online misinformation. Observing that the user comments on social media posts routinely include (mis)information that differs from the actual linked content, the authors devise a survey experiment on MTurk (n=954) in which participants either: (1) read a full article about Trump's approval rating; (2) read an article preview as it would appear on Facebook; (3) read an article preview with comments attached suggesting Trump's approval was higher than the article reported; and (4) read an article preview with comments suggesting Trump's approval was lower than reported. They find that respondents assigned to treatments with misinformed commentary (wrongly stating Trumps approval was higher/lower than that in the linked article) were significantly more misinformed about the topic. As a result, the use of social media for news—which introduces a new dynamic of attached user commentary—can contribute to a heightened vulnerability to misinformation among the U.S. public.

Finally, Avram et al (2020) contend that by making engagement metrics with news feed posts public (e.g., number of likes, shares), social media companies increase their users' vulnerability to online misinformation. Their study involved a 19-month deployment of a news literacy game that simulated a social media feed that—over the course of running it—presented 8,500 users (mostly from the US) with approximately 120,000 news articles (60,000 of which were from lowcredibility sources). Their findings show that posts displaying high engagement metrics (e.g., many likes/shares) strongly and significantly influenced the likelihood that users interacted with lowcredibility information while also reducing the likelihood that users fact-checked that information (Avram et al 2020, 1). Accordingly, misinformation posts that achieve high-engagement are more likely to cascade further online because social media users treat engagement metrics as a cognitive shortcut for assessing credibility.

4.4 Conclusion

In reviewing research on the factors contributing to heightened vulnerability to misinformation in the United States, this section has identified three main factors to consider when assessing the nature of the problem. First, partisan polarization (of the public, media, and political elites) creates an environment in which individuals are more likely to believe and spread ideologically congruent misinformation—a dynamic that has been found to be especially strong among right-leaning Americans. Second, cognitive states such as anger, anxiety, and highly emotional reasoning can prime individuals to be less discerning in information consumption and more receptive to misinformation— with some studies showing an interaction between emotional states and partisanship. Third, the increased reliance on social media for news consumption and accessing information on current events introduces new dynamics into American's information diet that likely gives misinformation more of an edge than in the past—the proliferation of low-credibility content, filter bubbles and echo chambers, and fallacious cognitive shortcuts (comment threads, engagement metrics)—which increase the consumption of political misinformation in the modern period.

References

[1] Allcott, Hunt, and Matthew Gentzkow. (2017). “Social Media and Fake News in the 2016 Election.” *Journal of Economic Perspectives* 31(2): 211–236.

[2] Anspach, Nicolas M., and Taylor N. Carlson. (2020). “What to Believe? Social Media Commentary and Belief in Misinformation.” *Political Behavior* 42(3): 697–718.

[3] Avram, Mihai, Nicholas Micallef, Sameer Patil, and Filippo Menczer. (2020). “Exposure to Social Engagement Metrics Increases Vulnerability to Misinformation.” *arXiv Preprint*: 1–9.

[4] Badawy, Adam, Emilio Ferrara, and Kristina Lerman. (2018). “Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign.” *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*: 258–265.

[5] Badawy, Adam, Kristina Lerman, and Emilio Ferrara. (2019). “Who Falls for Online Political Manipulation?” *Companion Proceedings of The 2019 World Wide Web Conference*: 162–168.

[6] Benkler, Yochai, Robert Faris, and Hal Roberts. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics.* New York: Oxford University Press.

[7] Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber. (2013). “Partisan Bias in Factual Beliefs about Politics.” *National Bureau of Economic Research*.

[8] Ferrara, Emilio. (2017). “Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election.” *arXiv Preprint*: 1–33.

[9] Flaxman, Seth, Sharad Goel, and Justin M. Rao. (2016). “Filter Bubbles, Echo Chambers, and Online News Consumption.” *Public Opinion Quarterly* 80(S1): 298–320.

[10] Freiling, Isabelle, Nicole M. Krause, Dietram A. Scheufele, and Dominique Brossard. (2021). “Believing and Sharing Misinformation, Fact-Checks, and Accurate Information on Social Media: The Role of Anxiety During COVID-19.” *New Media & Society*: 1–22.

[11] Greenstein, Michael, and Nancy Franklin. (2020). “Anger Increases Susceptibility to Misinformation.” *Experimental Psychology.*

[12] Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. (2019). “Fake News on Twitter During the 2016 US Presidential Election.” *Science* 363(6425): 374–378.

[13] Guess, Andrew, Brendan Nyhan, and Jason Reifler. (2018). "Selective Exposure to Misinformation: Evidence from Consumption of Fake News During the 2016 US Presidential Campaign." *European Research Council* 9(3): 1–14.

[14] Havey, Nicholas Francis. (2020). "Partisan Public Health: How Does Political Ideology Influence Support for COVID-19 Related Misinformation?" *Journal of Computational Social Science* 3(2): 319–342.

[15] Hornsey, Matthew J., Matthew Finlayson, Gabrielle Chatwood, and Christopher T. Begeny. (2020). "Donald Trump and Vaccination: The Effect of Political Identity, Conspiracist Ideation and Presidential Tweets on Vaccine Hesitancy." *Journal of Experimental Psychology* 88: 1–8.

[16] Humprecht, Edda. (2019). "Where 'Fake News' Flourishes: A Comparison Across Four Western Democracies." *Information, Communication, & Society* 22(13): 1973–1988.

[17] Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. (2000). "Misinformation and the Currency of Democratic Citizenship." *The Journal of Politics* 62(3): 790–816.

[18] Meirick, Patrick C. (2013). "Motivated Misperception? Party, Education, Partisan News, and Belief in 'Death Panels'." *Journalism & Mass Communication* 90(1): 39–57.

[19] Moravec, Patricia, Randall Minas, and Alan R. Dennis. (2018). "Fake News on Social Media: People Believe What they Want to Believe When It Makes No Sense At All." *Kelley School of Business Research Paper* : 18–87.

[20] Nikolov, Dimitar, Alessandro Flammini, and Filippo Menczer. (2020). "Right and Left, Partisanship Predicts (Asymmetric) Vulnerability to Misinformation." *arXiv Preprint*: 1–13.

[21] Nyhan, Brendan. (2010). "Why the 'Death Panel' Myth Wouldn't Die: Misinformation in the Health Care Reform Debate." *Politics* 8(1): 1–24.

[22] Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J. Wood. (2020). "Taking FactChecks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability." *Political Behavior* 42(3): 939–960.

[23] Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Micael Bang Petersen. (2020). "Partisan Polarization is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter." *American Political Science Review* 115(3): 999–1015.

[24] Peterson, Erik, and Shanto Iyengar. (2021). "Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading?" *American Journal of Political Science* 65(1): 133–147.

[25] Ribeiro, Manoel Horta, Pedro H. Calais, Virg´ılio A. F. Almeida, Wagner Meira Jr. (2017). " 'Everything I Disagree with is #FakeNews': Correlating Political Polarization and Spread of Misinformation ." *Proceedings of Data Science + Journalism, Halifax, Canada*: 1–8.

[26] Schaffner, Brian F., and Cameron Roche. (2016). "Misinformation and Motivated Reasoning: Responses to Economic News in a Politicized Environment." *Public Opinion Quarterly* 81(1): 86–110.

[27] Shin, Jieun, and Kjerstin Thorson. (2017). "Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media." *Journal of Communication* 67(2): 233–255.

[28] Shrestha, Anu, Francesca Spezzano, and Indhumathi Gurunathan. (2020). "Multi-Modal Analysis of Misleading Political News." *Multidisciplinary International Symposium on Disinformation in Open Online Media*: 261–276.

[29] Swire, Briony, Adam J. Berinsky, Stephan Lewandowsky, and Ullrich KH Ecker. (2017). "Processing Political Misinformation: Comprehending the Trump Phenomenon." *Royal Society Open Science* 4(3): 1–21.

[30] Swire-Thompson, Briony, Ullrich KH Ecker, Stephan Lewandowsky, and Adam J. Berinsky. (2020). "They Might Be a Liar But They're My Liar: Source Evaluation and the Prevalence of Misinformation." *Political Psychology* 41(1): 21–34.

[31] Tornberg, Petter. (2018). "Echo Chambers and Viral Misinformation: Modeling Fake News as Complex Contagion." *PLoS One* 13(9): 1–21.

Tornberg, Petter. "Echo chambers and viral misinformation: Modeling fake news as complex¨ contagion." PLoS one 13, no. 9 (2018): e0203958.

[32] Uscinski, Joseph E., Casey Klofstad, and Matthew D. Atkinson. (2016). "What Drives Conspiratorial Beliefs? The Role of Informational Cues and Predispositions." *Political Research Quarterly* 69(1): 57–71.

[33] Valentino, Nicholas A., Vincent L. Hutchings, Antoine J. Banks, and Anne K. Davis. (2008). "Is a Worried Citizen a Good Citizen? Emotions, Political Information Seeking, and Learning via the Internet." *Political Psychology* 29(2): 247–273.

[34] Weeks, Brian E. (2015). "Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation." *Journal of Communication* 65(4): 699–719.

5. Models for content moderation

The following section provides a more opinionated overview of the literature surrounding the question "**what is the best model for content moderation?**" which, for the most part can be categorized into a) automated (AI) moderation; b) human moderation (CCM work); and c) volunteer moderation. It concludes by gauging the scholarly consensus (or lack thereof) on the question, and potential avenues for future research.

Prepared primarily by: LB

Prior to the COVID-19 Pandemic, scholars and tech companies alike were debating whether, how, and to what degree artificial intelligence (AI) could/should be used to bolster online content moderation practices. Over the last decade as communication platforms spread throughout the world and the quantity of user content ballooned, some scholars began to ask whether platforms even had a choice in utilizing AI; commercial content moderation (CCM) workers were already notoriously overworked, and some even suffer from PTSD and other enduring cognitive health effects. When the Pandemic hit, social media giants like Facebook, Google, YouTube, and Twitter sent workers home, leaning heavily on algorithms in their place, rendering the scholarly debate surrounding models of moderation as pertinent as ever. Even as a myriad of scholars are tackling the question of AI, the question of whether or not it might provide a better model for content moderation—both ethically and logistically—remains open.

However, before AI became the central tenet of the literature that it is today, scholars such as Sarah T. Roberts focused on what she terms "commercial content moderation (CCM)" workers and firms, whose services platforms use to screen content. In several articles **(Roberts 2014; 2016; 2016; 2017)** and a subsequent book **(Roberts 2019)**, she builds on work from **Gillespie (2010; 2015; 2018)** addressing the politics and policies of platforms by exploring how the unseen efforts of CCM workers affect how users experience social media and user-generated content and the effects of such work on moderators themselves.[1] She employs a *qualitative, ethnographic method centered around interviews with moderators, foreign and domestic*. She

[1] Gillespie's work—including his frequently cited book *Custodians of the Internet*—is generally qualitative, relying on news, the published guidelines and policies of platforms, anecdotal examples, and case studies to provide comprehensive histories of content moderation, identify problems, and offer suggestions for improvement.

shows that CCM work is largely outsourced—in a manner remarkably similar to call-centers—and that tasks associated with the work range from mind-numbingly repetitive at best, psychologically damaging at worst, and underpaid in any case.

Scholars have explored the regulation of digital content since the early years of the Internet and social media (e.g. **Dutton 1996**; **Kollock and Smith 1996**; **Lampe and Resnick 2004**; **Verhulst 2006)**, and Roberts' descriptions of the dirty work of digital laborers added fuel to existing interest regarding how to improve online content moderation, both at the user and moderator level. Roberts' work and additional literature surrounding CCM and its effects on works have helped engender nuance in the contemporary debate surrounding automated moderation, demonstrating that even before debating the ethics AI, there are serious challenges in human moderation worth grappling with as well. Apart from necessity due to scale and COVID-19, the ethical and logistical challenges associated with human-driven moderation Roberts and others illuminated have perhaps inadvertently provided the most cogent argument in favor of automated moderation.

Since, several scholars have delved into the weeds of how models of content moderation work—theoretically and in practice. **Klonick (2017)** broke up types of CCM work into *ex ante*, *ex post proactive*, and *ex post reactive* categories; **Chancellor et al (2016)** studied the effectiveness of moderation by applying a *mixed-methods approach* to a case study of pro-eating disorder communities on Instagram[2]; **Gerrard (2018)** used this same case in order to demonstrate how users can circumvent moderation efforts, albeit with less clear methods; **Gibson (2019**) analyzes ~250K anonymous comments over a 14-month period from two Reddit forums matched in topic and size, but with differing moderation policies of "safe space" and "free space," finding that the former spurred more moderator action and self-censorship, and fostered more positive discussions.[3] **Quodling (2019)** addressed how platform operators attempt

[2] **Chancellor et al (2016)** analyze lexical, behavior, and topical changes associated with the emergence of lexical variation in pro-ED communities. They offer a set of general rules to define lexical variants, and categorize them as root tags and tag chains, the set of all lexical variants of a given tag. They used Instagram's official API to collect >8M public posts in the pro-ED space; to account for posts including banned tags, which are not returned in the API, they sample for pro-ED tags that co-occurred with banned tags and create a candidate pro-ED post set.

[3] **Gibson (2019)** compiled a data set of ~300K comments made on two subreddits and analyzed their language using Linguistic Inquiry and Word Count (LIWC2015), which sorts and displays the words present in text into a variety of categories (e.g. linguistic, social/affect, and cognitive)

to govern users through policy and enforcement strategies by employing *ethnographic, interview-based case studies of Facebook, Google+, and Twitter users*; **Gillespie (2015)** investigated how much users actually understand about the rules for content and the processes for its moderation imposed by platforms; **West (2018)** asked using a *survey method* how content moderation systems shape the affective relationship between users and platforms; and a myriad of scholars (**Zarsky 2014**; **Gillespie 2017**; **Jørgensen 2018**; **Gorwa 2019**) have offered suggestions on how regulatory and legal frameworks might be imposed upon platforms to better set parameters for online speech.

However, not all platforms employ(ed) human content moderators and/or AI to moderate content, particularly in the early years of social media. Some use(d) volunteer "admin" style of moderation, a more primitive model which itself has a fairly extensive literature behind it. The majority of such works employ Reddit—composed of "subreddits," each of which has their own volunteer moderator(s) tasked with custodial content duties—as a case study, while others employ the likes of Wikipedia and Discord. Volunteer moderators create, support, and control public discourse for millions of people for no compensation, leading **Matias (2019)** to ask, "what is the meaning of this work, and who is it for?" Using a *mixed-methods approach* based around an analysis from over 52,000 subreddits and interviews, he ultimately likened volunteer moderation to a form of "civic labor" akin to journalists who fact-check news on Facebook.[4] A decade earlier, **Postigo (2009)** asked a similar question with respect to AOL's extensive use of volunteers and the Dept. of Labor investigation which followed, *arguing based on interviews and DOL documents (qualitatively)* that the success of a co-productive relationship between platform and volunteer exists as a function of a balance between 1) the perceived reasonable compensation of volunteers, 2) social factors such as sense of community, and 3) a sense of accomplishment. **Squirrell (2019)** also addressed this question, employing an *ethnographic approach* to analyze two subreddits; he determined that moderators of forums with an epistemic

by frequency. The posts are then filtered at the author and comment levels for crossposters. They then conduct simple linear regressions to model the rate at which each subreddit gained subscribers and constructed linear mixed models fit by REML for each linguistic dimension to minimize error.

[4] Matias employs a variety of methods and data sources: participant observation, content analysis, interviews, and trace data collection on Reddit. Despite his collection of the latter, however, there are not statistical tests performed in the article.

focus utilize the tools at their disposal to mediate trust and establish a paradigm of constructive discourse, deeming this phenomenon "platform dialectics."

But even as scholars produce generally encouraging findings in case studies of volunteer-moderated communities, none argue that this model of moderation should (or even could) be applied to the major platforms (Facebook, Twitter, Instagram, YouTube). The sheer quantity of content posted by users on the major platforms has become so voluminous that some (e.g. **Gillespie 2020**) suggest that even CCM work is insufficient to handle the scale/size.[5] In sum, though research on volunteer moderation is largely positive—it does little to answer the principle question of which model is best.

Before pivoting to AI and the crux of the contemporary debate around that question, a practice present in each model of moderation warrants acknowledgement in this literature review: flagging. Flagging describes the mechanisms by which users report offensive content to an online platform, serving as both a solution to the problem of scale and as a rhetorical justification for platform owners removing content. **Crawford and Gillespie (2016)** argue that, though simple in the abstract, in practice flagging is a complex and strategic interaction between users, algorithms, moderators, and platforms themselves. Unlike the predominately ethnographic works covered thus far, Unlike the predominately ethnographic and case study-based works on content moderation, **Naab et al (2018)** employed a controlled experiment and a subsequent regression model to test a research question (RQ1)— whether flagging behavior differs when a uncivil comment receive agreement, disagreement, or no response—and three hypotheses—that perceived self-responsibility mediates the effect of agreement or disagreement on flagging behavior (H1); that attribution of responsibility to professional moderators mediates the effect of agreement or disagreement on flagging behavior (H2); and that response politeness moderates the effect of response direction on perceived self-responsibility, attributed professional responsibility, and on flagging behavior (H3). They find that flagging was more likely when intervention information was presented; that providing intervention significantly increases flagging behavior in nearly all groups; and that there exists a significant indirect effect of response direction through attributed professional responsibility on flagging behavior as well as for the interaction of response direction and politeness. Ultimately, they conclude that

[5] Many scholars studying the subject refer to this "problem of scale," though in his 2020 article Gillespie adamantly argues this is incorrect and should be reconceptualized as a problem of size.

implementation of a flagging tool not only saves resources but integrates users in the negotiation of shared values.[6]

Not all research on flagging has been positive, though; **Mármol, Pérez, and Pérez (2014)** demonstrate that malicious users may report harmless content just to compromise other users' material and that the large number of users flagging content as offensive makes its management an even more difficult task for CCM workers, in turn proposing a novel reputation-based approach with which social networking sites are capable of automatically assessing accusers' honesty before withdrawing any content. In sum—though not a model itself—flagging is a crucial mechanism within each model of content moderation, and questions surrounding its effectiveness, how it should function procedurally, and whether it helps simplify CCM work or in fact further complicates the efforts of human moderators persist.

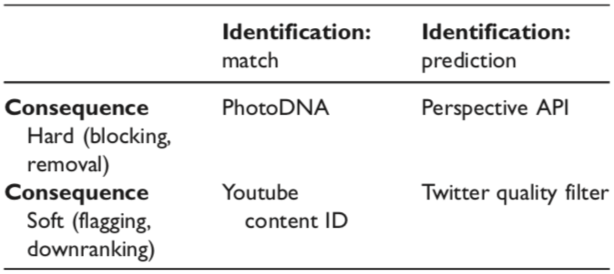
The rapid growth in AI technology over the past decade mirrors its swift rise to top of content moderation literature. Though some scholars, particularly computer scientists, have previously researched automation in content moderation—such as **Delort et al (2011)**, who *describe a novel classification technique to train a classifier from a partially labeled corpus and use it to moderate inappropriate content on online discussion sites, implement, and test it on a corpus of messages and compare it with two baseline techniques* well before automated moderation became the pertinent issue it is today; **Gehl, Moyer-Horner, and Yeo (2016)**, who *review peer-reviewed articles on computer vision-based pornography filtering (CVPF)* and concluded that the technology has generally trained computers to "see" a very specific, far too idealized form of pornography; **Binns et al (2017)**, who study the inheritance of bias in algorithmic content moderation by *training classifiers on comments labeled by different demographic subsets (men and women)* to test and understand how differences in conceptions of

[6] To answer RQ1, H1, and H2 (Study 1), Naab et al conduct a 2x2x3 between-subjects design, varying intervention information (available vs. unavailable), type of victim (individuals vs. social group), and type of response (agreement vs. disagreement vs. none). They then conduct treatment checks and construct a binary logistic regression model based on the results. To answer H3 (Study 2), they conduct a 2x2 between-subjects design, varying response direction and politeness of the response. To test if the manipulation for response direction was successful, they conduct an analysis of variance (ANOVA), and a second ANOVA with the experimental groups as independent variable to test for the perceived difference in response politeness. They then construct an indirect logistic regression model using PROCESS for SPSS, model 8.

offense between them might affect the performance of the resulting models on various test sets; **Gröndahl et al (2018)**, *who reproduce seven hate speech detection models*, determining that model architecture is less important than data type and labeling criteria; **Jhaver et al (2019)**, who *participated as Reddit moderators for over a year and conducted interviews with moderators* to understand the use and effectiveness of its configurable, automated program called "Automod*"*; **Munk (2017)**, who did an *opportunity/cost analysis of the false positives/false negatives associated with predictive counter-terrorism algorithms based on a review of literature* and concluded that the algorithms are ineffective, risky, and inappropriate (identifying ~100K false positives for every real terrorist); and a multitude of others (e.g. **Arango et al 2019**; **Mathew et al 2019**; **Vijayaraghavan et al 2021**; **Yin and Zubiaga 2021**) —while eminent scholars of content moderation (e.g. Gillespie) who focus on the theoretical aspects of moderation as opposed to the technical, have only recently begun to tackle the subject.

Tarleton **Gillespie (2020)**—perhaps the eminent scholar of this field—argued that, though AI seems like the perfect response to growing moderation challenges, we should be hesitant to automate. He does not conduct an experiment or propose hypotheses; rather, he articulates arguments against automation, and suggests a profound act of "countercapitalist imagination" to find new ways to fit machine learning techniques with new forms of governance may be necessary. **Gorwa, Binns, and Katzenbach (2020)** offer a more robust look at algorithmic content moderation, though, like Gillespie's, it *does not utilize any particular methods; rather, they provide a typology of moderation and an overview of the publicly reported algorithmic moderation systems deployed by major platforms*, replicated below:

Table 1. Simple typology of moderation, with examples.



	Identification: match	**Identification:** prediction
Consequence Hard (blocking, removal)	PhotoDNA	Perspective API
Consequence Soft (flagging, downranking)	Youtube content ID	Twitter quality filter

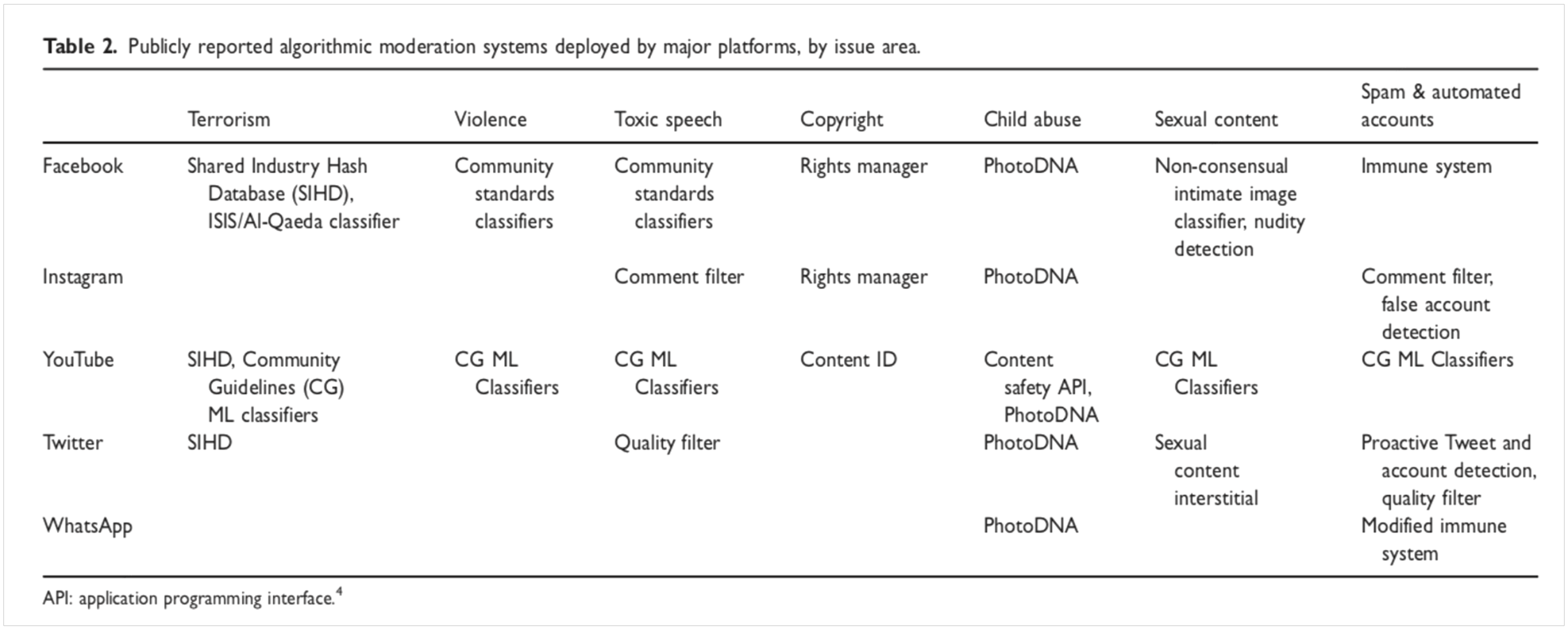


Table 2. Publicly reported algorithmic moderation systems deployed by major platforms, by issue area.

	Terrorism	Violence	Toxic speech	Copyright	Child abuse	Sexual content	Spam & automated accounts
Facebook	Shared Industry Hash Database (SIHD), ISIS/Al-Qaeda classifier	Community standards classifiers	Community standards classifiers	Rights manager	PhotoDNA	Non-consensual intimate image classifier, nudity detection	Immune system
Instagram			Comment filter	Rights manager	PhotoDNA		Comment filter, false account detection
YouTube	SIHD, Community Guidelines (CG) ML classifiers	CG ML Classifiers	CG ML Classifiers	Content ID	Content safety API, PhotoDNA	CG ML Classifiers	CG ML Classifiers
Twitter	SIHD		Quality filter		PhotoDNA	Sexual content interstitial	Proactive Tweet and account detection, quality filter
WhatsApp					PhotoDNA		Modified immune system

API: application programming interface.[4]

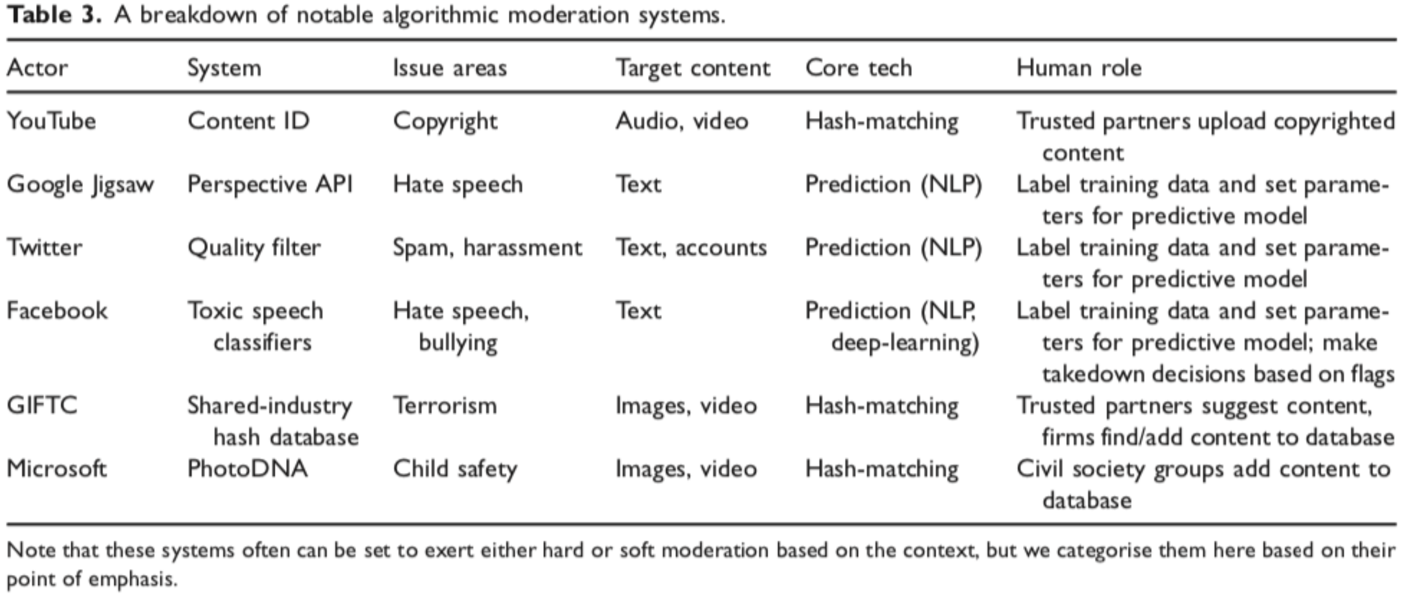


Table 3. A breakdown of notable algorithmic moderation systems.

Actor	System	Issue areas	Target content	Core tech	Human role
YouTube	Content ID	Copyright	Audio, video	Hash-matching	Trusted partners upload copyrighted content
Google Jigsaw	Perspective API	Hate speech	Text	Prediction (NLP)	Label training data and set parameters for predictive model
Twitter	Quality filter	Spam, harassment	Text, accounts	Prediction (NLP)	Label training data and set parameters for predictive model
Facebook	Toxic speech classifiers	Hate speech, bullying	Text	Prediction (NLP, deep-learning)	Label training data and set parameters for predictive model; make takedown decisions based on flags
GIFTC	Shared-industry hash database	Terrorism	Images, video	Hash-matching	Trusted partners suggest content, firms find/add content to database
Microsoft	PhotoDNA	Child safety	Images, video	Hash-matching	Civil society groups add content to database

Note that these systems often can be set to exert either hard or soft moderation based on the context, but we categorise them here based on their point of emphasis.

Like Gillespie, they conclude that even "well-optimized" moderation systems could exacerbate many existing problems by further increasing opacity; further complicating outstanding issues of fairness and justice; and re-obscuring the political nature of speech decisions being executed at scale. **Cobbe (2020)** also offers a substantive evaluation of AI in content moderation, similarly *without employing methods or statistical analysis*. She similarly argues the emergence of algorithmic censorship is an undesirable development, by permitting platforms to more effectively align public and private online communications while simultaneously undermining those platforms' function as a space for communication. **Ruckenstein and Turunen (2019)** argue that the current platform logic forces moderators to operate like machines themselves and maintain that the work of CCM should be reframed to integrate moderators' aims with AI systems; however, they don't offer much with respect to how that might look in practice, and *nothing by way of statistical/data testing*. In any case, it is apparent that scholarship on AI and content moderation has only very recently begun to take off, and there is a substantial gap in the literature regarding methodological analyses of the subject.

The closest article I've identified to answering the principle question at hand comes from **Einwiller and Kim (2020)**, who conduct two studies in the U.S., Germany, South Korea, and China to examine how platforms approach moderation—one employing content analysis of policy documents, and another utilizing interviews with representatives of online content providers. They find that content guidelines are often communicated in user-unfriendly terms of service, while only Korean platforms present their policies vividly, and that platforms are not proactive enough in communication for preventing harmful content, often focusing more on avoiding legal ramifications than on educating users. However, Einwiller and Kim (2020) primarily seek to answer what method of *communicating* content moderation policies to users is best, rather than the model of moderation itself (i.e. AI vs. CCM vs. volunteer).

In conclusion, there appears to be no clear scholarly consensus on how platforms should moderate content in a broad sense; however, scholars do seem to agree that AI may in fact prove a cure worse than the disease, and that we should be hesitant to automate the moderation process.

6. Strategies for content moderation and their effect on malicious content

The following section provides a more opinionated overview of the literature.
Prepared primarily by LB

Background

As addressed in the previous section, a substantial chunk of scholarship on online content moderation concerns the strategies and methods employed by platforms to remove hateful, obscene, or otherwise malicious content. Often, scholars have utilized one or more case studies of platforms and/or aggregate data released by platforms themselves (see pp. 2-3) to gauge the effectiveness of such methods. The following section amalgamates and summarizes several dozen of these articles, their methods (elaborated upon in footnotes where applicable/quantitative), and their ultimate conclusions, providing a general overview of the most common methods of moderation and the challenges faced in implementing them. I've also added a full bibliography at the end, which includes links to access each article referenced. Sections are ordered as I wrote them and for smooth transitions, and in no way reflect the breadth of research on the subject.

In the broadest sense, content moderation strategies can be divided into two main approaches: one reliant upon extensive regulation of user-generated content, and one that mostly leaves this task to the community itself; this division can be conceptualized the same way as AI and CCM moderation versus volunteer moderation mentioned in the previous section. **Artime et al (2020)** recently compiled a massive dataset comparing Twitter (moderated) and Gab (unmoderated), using keywords and sentiment analysis to reconstruct behavioral networks on online social platforms. Their methodology is impressive, and their finding unexpected—moderation approach does not appear to be a crucial factor to consider in the assessment of the effectiveness of network dismantling. However, since the present section concerns strategies *for* moderating content (as well as the fact that few platforms—and none of the major ones—actually employ a volunteer-centric approach), this section focuses on the array of methods used by platforms which themselves regulate user content.

The most prominent/common of these methods are **content removal, user bans, hashtag bans, automated filtering** (e.g. image/video) and **algorithmic sorting, content restriction** (e.g. 18+, NSFW)**, and content labeling** (e.g. misinformation warnings). Given that major platforms recently began implementing or ramping up content labeling—most notably in response to COVID-19 and 2020 Election misinformation—I personally believe the question of labels, which I address last, provides particularly fertile ground for new research. Another promising ground for future research is the growing use of artificial/automated intelligence (AI) to off-source and/or augment the labor associated with the aforementioned methods.

At the outset, it is worth highlighting the great obstacle obstructing scholars seeking to understand how platforms moderate and how effective those strategies are—platforms themselves. As eminent content moderation scholar Sarah T. Roberts noted, "platforms operate their moderation practices under a complex web of nebulous rules and procedural opacity" and "are extremely reluctant to acknowledge, much less describe in great detail the internal policies and practices that govern what goes into the content ultimately available on the mainstream site. **(Roberts 2018)**. Many platforms do publish periodic transparency reports—which typically disclose aggregate, quantitative data overviews about requests for content removal, rates of compliance on the platform, and sometimes the frequency of user appeal—but these have major limitations, even for more thorough reports such as the Community Guidelines enforcement reports published by Facebook, Twitter, and YouTube for the first time in 2018. For one, the aggregated data only shows the platforms' assessments, not the merits, inhibiting researchers from evaluating the accuracy of decisions. Further, transparency reports vary widely from company to company with respect to how they classify data and how much detail they provide.[1] Accordingly, researchers needing to know what content was actually removed often turn to the Lumen database, which archives legal takedown notices from any platform or sender that chooses to share them; to date, it is the best source for such information, holding some 9.3M notices targeting approximately 3.35B URLs, mostly form Google.[2] Scholars have also turned to issue-specific platform disclosures, such as Google's 2018 report *Three Years of the Right to Be Forgotten*, as well as public filings, other government disclosures, and published reports from independent auditors such as The Global Network Initiative **(Keller and Leerssen 2020)**. **Anderson, Stender, Myers West, and York (2016)** turned to crowdsourced/solicited data

[1] For a comprehensive overview of transparency reporting, see **Singh and Bankston (2018),** who issued a report on the practice on behalf of New America's Open Technology Institute.

gathered over the course of four months from a questionnaire on the onlinecensorship.org website, aggregating and analyzing the data across geography, platform, content type, and issue areas to highlight trends in social media censorship and variation between platforms. In other words, aggregate, comparable data on platform moderation decisions can be so difficult to come by that even highly respected scholars have thrown major selection bias concerns to the wind just to have *something* to work with.

Though limitations seem to have driven some scholars to write theoretically or qualitatively on content moderation by platforms, some have nevertheless endeavored to methodologically test hypotheses they've posed using data they've been able to access, often focusing on a single platform as a case study. As Gillespie pointed out, despite the fact that research in the field tends to be driven by high-profile incidents and people (e.g. misinformation following the 2016 Election; domestic terrorism and online hate following the Christchurch shooting), "as concern about moderation grows, scholarly attention has grown with it—somewhat—from specific controversies to deeper, structural questions about how moderation is organized and enacted" **(Gillespie et al 2020).**[3]

Content Removal & Account Bans

Though content takedowns are only one way platforms moderate, a breadth of research topics could conceivably fall under that umbrella. In a literature review, Keller and Leerssen categorize current and future empirical research about platform content takedowns as follows: accuracy rates in identifying prohibited material; areas of higher or lower accuracy; success rates of mechanisms designed to prevent over-removal; costs; filters; community guidelines; and consequences of removal, over-removal, and under-removal **(Keller and Leerssen 2020)**.

Removal is, of course, the most common and simple way to moderate content. Facebook, Instagram, Twitter, YouTube, Reddit, Tumblr, Snapchat, Russia's VK, China's WeChat, and virtually every other platform hosting user-generated content does this to some degree—adopting, enforcing, revising, and often avoiding what scholars increasingly refer to as platform

[3] I recommend reading the article cited here: "Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates." Co-written by nine eminent scholars of critical internet studies and published in *Internet Policy Review* Volume 9, Issue 4 (October 2020), the article proposes how to expand present research on content moderation.

governance or platform law **(Kaye 2019).** This concept, "a scholarly framework intended to capture the layers of governance relationships structuring interactions between key parties in today's platform society" **(Gorwa 2019)**, has received much theoretical attention as of late—particularly from **Gillespie (2010; 2015; 2018; 2018a)**, who set the stage for work researching how platforms "intervene" in everyday life, shape online experiences, and determine what information to make (in)visible, culminating in his comprehensive look at platforms as "custodians" of public speech through the work of moderation that they perform, *Custodians of the Internet*.[4]

Such a complex, dynamic theoretical framework is made necessary in large part because policing and the power of removal alone have proven wholly insufficient at combatting unwanted content, as new challenges and crises necessitate dynamic approaches. For example, ISIS fighters created the "Syrian Archive," a software that would download new material from Syrian YouTube channels to its servers before they were purged from the site, forcing the platform to ramp up moderation to an extent at which it could delete videos at a speed faster than the Archive could download them. A challenge of this scale and urgency necessitated the use of AI, which brings about its own problems, such as mistaken deletions or blockings, or ones that activists or governments disagree with. Scholars such as **Banchik (2020)** have brought this practice into question, as removal of potential human-rights violations and war crimes before detection can serve to destroy evidence. The YouTube "scraper" was subsequently implemented in December 2016; every ten seconds, the scraper builds a sample of 50 videos which were published in a ten second window one-minute prior.

In a study of the scraper and similar, large-scale content moderation systems, **Suzor (2020)** found that—assuming removed videos are close to average duration—the sample of videos tested is likely close to 10% of total videos uploaded to the platform, and of those, approximately 6.7% have since been removed because of some form of content moderation. YouTube also moderates content by restricting it, but not outright removing it. In such cases,

[4] Much like the difficulties described on pp. 2, the major research gap presently in the subfield of platform governance is the lack of systemic analysis of governance-in-practice in a form that enables comparison across contexts and over time. Though this section does not address platform governance any further, recent research is voluminous enough to warrant its own section. For more on platform governance, I recommend **Klonick (2017), Gorwa (2019),** and **Suzor et al (2018)**.

users must have an account to watch the video and overcome age restrictions. Suzor, however, finds that only ~0.3% of 90M videos scraped during the time period studies were issued such *content warnings*—another way platforms moderate content—far fewer than the number removed for terms of service violations, copyright violations, or listed as unavailable without further information.[5] In the same article, Suzor (2020) also looks at Twitter's tester, which has been running stably since January 2018. The Twitter tester follows the same logic as the YouTube tester, collecting a random sample of tweets as close as possible to the time they were posted. He argues that the error results that Twitter returns when it moderates content are, compared to YouTube, very basic; whereas YouTube data allow for categorical analysis of alleged malfeasance, when Twitter removes a tweet it typically applies a sanction to the user's account which will only be lifted after the offending tweet is deleted or appeal is won. Accordingly, scholars cannot easily distinguish between Tweets deleted because they violate Twitter's rules from tweets deleted by the user for other reasons. Suzor's article also studies Instagram's most visible and controversial mechanism to moderate content— which is expanded upon in the following section *(Hashtag Bans)*—as well as a more traditionally computer science-oriented article with Gray **(Gray and Suzor 2020)**, which employs a dataset of ~77M YouTube videos to explore how digital and computational methods can be improved and leveraged in large-scale moderation systems.[6]

One reason removing content alone is insufficient stems from the fact that the creator can simply continue their work, perhaps unchecked in future instances; accordingly, platforms have at times turned to *de-platforming*, generally in the form of account *bans* (e.g. Donald Trump). Even the most stringently-enforced account bans, however, have at times proved ineffective, or

[5] Suzor collects random samples of pubic social media content, and then tests the availability of that sample at a later date; he attempts to assemble the initial sample as close as possible to the time a post is uploaded or made public, test availability again between 8-20 days later, and then re-test it two months after it was initially collected (to identify posts that were reinstated). However, the article only reports on the content that was a) captured within an hour of being posted and b) was tested again between 8-20 days later.

[6] Gray and Suzor use the BERT language model to train a machine learning classifier to identify videos in categories that reflect ongoing controversies in copyright takedowns. They find that a simple neural network infrastructure, a pre-trained BERT language model, and substantial cloud processing power are able to achieve satisfactory performance on a multiclass classifier over short texts with a relatively small number of training examples.

at least insufficient. **Jardine (2019)** outlines what he calls the "surface-to-Dark Web content cycle," a four step process in which content is produced on the "surface web," is then restricted on the surface web, is displaced to the Dark Web, from where it percolates back to the surface web. He also outlines two easier pathways by which malicious content can resurface: 1) through direct resurgence on the surface web (e.g. *The Daily Stormer* trying to find a foreign domain to host its content after being delisted) and 2) by way of multi-media content pieces (e.g. memes, infographics, and slogans). The frequency at which this occurs is unknown—particularly percolation through idea networks; however, Jardine demonstrates that malicious content can make its way back onto platforms no matter how aggressively said platform retaliates against the offensive post and/or offender. Though the bulk of scholarship on content removal alone focuses on the insufficiencies of doing so, some work does attest to its impact. **Srinivasan et al (2019)** study content removal as a moderation strategy by leveraging a new quasi-experimental design they call "delayed feedback" upon a case study of the "ChangeMyView" subreddit, showing individuals subject to moderators (comment removal) in turn regulate their bad behavior, reduce toxicity in language, and reduce noncompliance, identifying a clear causal effect in the latter case.[7]

Another article which employs Reddit as a case study from **Jhaver et al (2019)** seeks to understand how the action of content removal affects the individuals who produced it. They conducted a survey of over 900 Redditors, finding that, although 18% accepted that their post removal was appropriate, the vast majority did not—37% did not understand, and 29% expressed some level of frustration about the removal.[8] Though their regression analyses do not (in my

[7] To identify the causal effects of moderation in the case study and account for the inability of standard interrupted time-series (ITS) analysis to prove that a user, after making a problematic comment, would have behaved worse if the comment had not been removed, Srinivasan et al leverage their "delayed feedback" approach. This exploits the observation that *some affected individuals make comments in the (often hours-long) "pre-removal" window wherein the to-be-removed comment has not yet actually been removed.*

[8] Jhaver et al conduct regression analyses on their DVs, 'fairness' and 'post again', adding control variables and posting context variables. They find that reading the rules was significantly associated with fairness even after controlling for demographics and prior Reddit history variables, but in the opposite direction expected; when users read the rules, they are less likely to consider the removal fair. They also find that when users perceive the rules to be clear they are more likely to see the removal as fair and more likely to post again.

view) uncover any dramatic findings, the question of how moderation—namely content removal—affects users' experiences and future behavior is a pertinent one to this discussion.

Though platforms' efforts to curb misinformation, hate speech, and graphic content dominate scholarship on removal, there are other reasons for removal worth considering, such as copyright violation. Legal scholars **Urban et al (2016)** wrote *Notice and Takedown in Everyday Practice*, a review of a sample of around 3,500 takedown request vis-à-vis the Digital Millennium Copyright Act's notice and takedown process (Section 512), which they argue may cause harm to users' expression through targeting and takedown of non-infringing content. They specifically find that 31% of takedown requests appear to contain mistakes of various sorts.[9] In a review of this "herculean task," however, **Ford (2017)** calls the project "completely unpersuasive," arguing that no evidence of consequential damage to expressive rights is presented by the authors. He argues that a higher level of mandated care to avoid false flags as recommended by Urban et al may result in an improvement in accuracy but also a reduction of search, and, consequently, an increase in infringement.[10] This is in line with **Engstrom and Feamster (2017)**'s argument in a report on the DMCA as it pertains to automated filtering/content detection tools; they conclude that calls to require OSPs to implement technical measures to police infringement overestimate the technical capacity of such tools while underestimating the damage that mandating their use would have upon the internet ecosystem.

Hashtag Bans

Challenges in moderating harmful online content often extend beyond individuals, and even specific hate groups, sometimes to entire "communities." To supplement human-led and

[9] Urban et al conduct three empirical studies: Study 1 is a qualitative, survey-driven documentation of the ways in which the notice and takedown process have been perceived and operationalized by major U.S. online service providers and rightsholders; Study 2 quantitatively examines a random sample of takedown notices, taken from a set of over 108M requests submitted to the Lumen archive over a six-month period; and Study 3 provides a further quantitative examination of a random sample of notices that were sent to Google in relation to its Google Image search from the same six-month set of takedown requests. These quantitative studies mainly involve summary statistics categorically arranged into visualizations, but not statistical testing.

[10] The bulk of Ford's critiques come from methodological grounds (e.g. improper extrapolation of non-representative data, unpublicized data, improper interpretation of findings, etc.)

AI-driven attempts to screen and remove malicious content, some platforms have attempted to curb malicious content by blocking entire *hashtags*, essentially attempting to rectify these insufficiencies by containing the spread of potentially harmful content into the rest of the platform and preventing the formation of communities around harmful topics. This strategy is most associated with Instagram, on which hashtags are the predominant method of searching for content from Instagram users that a person is not already following. Instagram moderates hashtags in four different ways: 1) a 'hard' block approach, in which a user visiting these tags will encounter an HTTP 404 error; 2) only displaying a selection of 'top posts' of a hashtag; 3) displaying an interstitial 'content warning' for tags related to eating disorders, self-harm, firearms, animal abuse, and at times providing links to third-party helplines; and 4) a 'silent' block approach, in which the platform presents a false statement that there are "no posts yet" on the tag.

Unlike YouTube and Twitter, Instagram provides no public application programming interface (API) that can be used to build a random sample of posts nor does it provide any public information about whether it has moderated a post, and there is little easily discernible pattern that explains why it chooses some forms of blocking over others. Moreover, the status of hashtags sometimes changes over time **Suzor (2020)**. This renders researching hashtag moderation by Instagram even more difficult than researching YouTube's scraper or Twitter's tester, but Suzor (2020) and The Observatory nonetheless track it, as well as how Instagram moderates posts (see footnote).[11]

Duguay et al (2020) described Instagram's hashtag moderation as 'patchwork platform governance', an approach that relies on formal policies and content moderation mechanisms but pays little attention to dominant platform technocultures and their sustaining architecture that they found applicable to Tinder, Instagram, and Vine. Using the close, qualitative lens of queer women's experiences, they find that formal governance measures like Terms of Service and

[11] In order to track which hashtags are blocked on Instagram, Suzor (and The Observatory) first developed a super-set of hashtags that had been reported as blocked by news outlets and supplemented it with search results on a major search engine for the error messages that Instagram displays when it blocks a hashtag or hides it behind an interstitial content warning. They then check these 1,449 hashtags daily for availability and separately test 5000 hashtags that have been most frequently used with posts that use previously-blocked hashtags in the sample. This provided ~12K tags tested each day, and an eventual sample of 12.6K hashtags that have been blocked at least once.

flagging mechanisms did not protect users from harassment, discrimination, and censorship and that key components of the platforms' architectures (e.g. hashtag bans, cross-platform connectivity, and algorithmic recommendation systems) reinforced this, arguing in turn that platforms should take more systematic approaches to governance.

Instagram and Pinterest were vocal about their decisions to police pro-eating-disorder content (uncharacteristically so, given that platforms regularly downplay intervention), announcing new, targeted policies through press releases **(Gillespie 2015)**. **Chancellor et al (2016)** tests the effectiveness of these strategies by conducting a case study on pro-ED communities on Instagram, compiling a dataset of 2.5M posts from 2011-2014. They find that, in response, the pro-ED community adopted non-standard lexical variation of moderated tags to circumvent these restrictions, and these variations grew increasingly complex over time (e.g. "#thyghgapp). Ultimately, they argue that despite Instagram's moderation strategies, pro-ED communities are active and thriving, demonstrating its insufficient effectiveness as an intervention for communities of deviant behavior.[12] **Gerrard (2018)** studied the same community—extending to sites like Pinterest and Tumblr and adopting a more qualitative approach—finding not only that untagged pro-ED content can easily be found and that users are evading hashtags, but also that platforms' algorithmic recommendation systems in fact recirculate pro-ED content, revealing the limitations of hashtag logics in social media content moderation. In any case, while hashtag moderation may be necessary, the overwhelming scholarly consensus at present appears to be that the governance process is often inconsistent and opaque—particularly with respect to Instagram—and that the results are insufficiently positive.

Anti-Extremism and Anti-Propaganda[13]

[12] Chancellor et al analyze lexical, behavior, and topical changes associated with the emergence of lexical variation in pro-ED communities. They offer a set of general rules to define lexical variants, and categorize them as root tags and tag chains, the set of all lexical variants of a given tag. They used Instagram's official API to collect >8M public posts in the pro-ED space; to account for posts including banned tags, which are not returned in the API, they sample for pro-ED tags that co-occurred with banned tags and create a candidate pro-ED post set.

[13] Though some can be found here, most work on combatting misinformation is summarized in the *Labeling* section.

An area of concern in content moderation worth very briefly addressing—though it intersects with strategies and effects and is not itself a method—has been on combatting propaganda and extremist messaging. The development of digital media and its decentralized mode of content production complicates the fundamental notion that disguised propaganda derives from large-scale organizations through one-to-many communications channels and complicates frameworks for identifying and analyzing it **(Farkas and Neumayer 2017; Farkas and Neumayer 2020)**. "Bot" user profiles have made this especially challenging, particularly considering the insufficient capacity at present for algorithms to adequately analyze the cultural contexts of each post. Thus, identifying hate speech and disguised propaganda has largely fallen on human content moderators (the limitations upon whom are outlined extensively in the previous section).

Relatedly, some scholars have researched how platforms make and justify decisions regarding permissible political advertising on their platforms. **Kreiss and McGregor (2019)** interviewed former social media firm employees and political practitioners and analyzed emails (both small *N*) exchanged between Facebook government and elections staffers and two campaigns regarding the platform's policies in the context of paid speech, concluding that an insufficient degree transparency presently exists with respect to how Facebook and Google make decisions and justify how their policies are applied and enforced, inhibiting the ability for political practitioners and the public to exercise their voice and/or contest regulation decisions. A year prior, they used a "soak and poke" style interview-based approach to look specifically at how Microsoft, Facebook, Twitter, and Google shaped political communication by working with campaigns during the 2016 U.S. Presidential Cycle. They find that technology firms are motivated to work in the political sphere for marketing, advertising, and relationship-building in the service of lobbying efforts, and that, to facilitate this, these firms have developed organizational structures and staffing patterns that align with the partisan nature of American politics. Further, they conclude that Facebook, Twitter, and Google in particular go beyond promoting their services and facilitating ad buys, and actively shape campaign communication through their close collaboration with political staffers **(Kreiss and McGregor 2018)**. Alas, given the impossibility of gathering aggregate data on the relationships between platforms and campaigns, there is little quantitative work exploring this subject.

Additionally, attention has recently been paid by scholar to the role that platforms like YouTube themselves play in radicalization, especially by way of their algorithmic recommendation system. **Munn (2019),** as well as prominent journalists **(Naughton 2018; Tufekci 2018)** have argued in a qualitative fashion that YouTube acts not just as a platform upon which extremist content can be uploaded, but a pathway and pipeline toward radicalization. **Ledwich and Zaitsev (2019)** refute this popular claim, as their quantitative analysis suggest that recommendation algorithms actively discourage radicalizing or extremist content, and instead favor mainstream media and cable news content.[14] In any case, debate over whether algorithmic recommendation systems steer users away from seedy content or in fact dig them deeper into the proverbial rabbit hole bespeaks broader questions on AI, its effectiveness at combatting unwanted content, and the role it should (or should not) play in the content moderation process.

Algorithmic Moderation

As with the last section, AI warrants its own discussion under the umbrella of strategies and effects. And, once again, the scholarly consensus appears to be on exercising caution, as the technology could prove a cure worse than the disease. Algorithmic and automated content moderation are perhaps the hottest topic within the literature at the moment, and several questions asked under that umbrella apply to the present section. Prior to its recent expansion—driven by the ever-growing size of platforms, the COVID-19 pandemic, and advancements in automated technology—most scholarly work on AI as it pertains to moderation was conducted by computer scientists.

For instance, in 2011**, Delort et al (2011)** described a novel classification technique they created to train a classifier from a partially labeled corpus to moderate inappropriate content on online discussion sites, implement it, and test it on sites against baseline techniques.[15] Another early use of AI in content moderation was implemented to screen pornography; computer vision-based pornography filtering (CVPF) is a subfield in computer science training computers on how to recognize the difference between pornographic and non-pornographic digital imagery. **Gehl,**

[14] Ledwich and Zaitsev categorize ~800 political channels on YouTube and then differentiate between political schemas to analyze the algorithm traffic flows out between each group.

[15] For articles such as this in fields which use methods entirely alien to me (i.e. computer science), I do not elaborate upon the methodological approach.

Moyer-Horner, and Yeo (2016), upon review of 102 peer-reviewed CVPF articles, determined that the technology as a whole trained computers to see a specific, idealized form of pornography (e.g. lone, thin, naked women), a finding **Saka (2020)** cites in her qualitative argument that algorithms are not necessarily value-neutral phenomena and have been shown to operate in a gender-biased manner.

AI's function in moderation has recently expanded far outside pornography and retroactive content removal (e.g. the aforementioned YouTube scraper), venturing further into other aspects of the moderation process and, accordingly, questions surrounding AI have branched ever increasingly out into scholarly realms apart from computer science. **Binns et al (2017)** studied the inheritance of bias in algorithmic content moderation by training classifiers on comments labeled by different demographic subsets (men and women) to test and understand how differences in conceptions of offense between them might affect the performance of the resulting models on various test sets; **Gröndahl et al (2018*)*** reproduce seven hate speech detection models, determining that model architecture is less important than data type and labeling criteria; **Jhaver et al (2019b)** participated as Reddit moderators for over a year and conducted interviews with moderators to understand the use and effectiveness of its configurable, automated program called "Automod*"*; **Munk (2017)**, conducted an opportunity/cost analysis of the false positives/false negatives associated with predictive counter-terrorism algorithms based on a review of literature and concluded that the algorithms are ineffective, risky, and inappropriate (identifying ~100K false positives for every real terrorist).

Another automated tool of moderation is textual analysis. In a report issued for the *Center for Democracy & Technology* on automated social media content analysis, **Duarte, Llanso, and Loup (2017)** argue that today's automated content analysis technologies have limited ability to parse the nuance of human communication or detect the intent/motivation of the poster. The five key limitations, they argue, are that: 1) natural language processing tools perform best when they are trained and applied in specific domains and cannot necessarily be applied with the same reliability across different contexts; 2) decisions based on automated social media content analysis risk further marginalizing and disproportionately censoring groups that already face discrimination; 3) natural language processing tools require clear, consistent definitions of the type of speech to be identified, and policy debates around content moderation and social media mining tend to lack such precise definitions; 4) the relatively low accuracy and intercoder reliability achieved in natural language processing studies warn strongly against widespread application of the tools to social media content moderation; and 5) even state-of-the-art natural language processing (NLP) tools remain easy to evade and fall short of humans' ability to parse meaning from text.

In the strongest recent article I've found on algorithmic content moderation, **Gorwa, Binns, and Katzenbach (2020)** break down the typology and function of various algorithmic moderation systems, which platforms employ them, and to what end, and the human role within these systems. rather than attempt to recount that large swatch of information here, I have pasted their tables below:

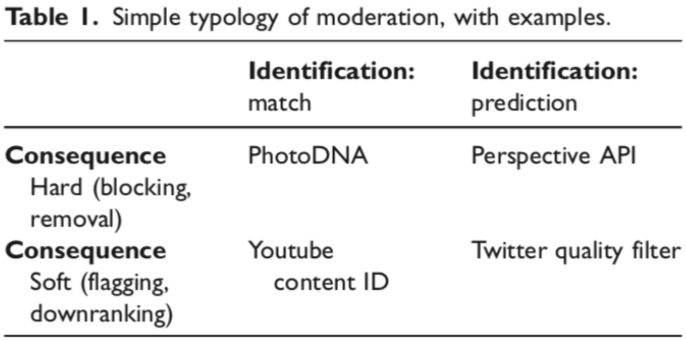


Table 1. Simple typology of moderation, with examples.

	Identification: match	**Identification:** prediction
Consequence Hard (blocking, removal)	PhotoDNA	Perspective API
Consequence Soft (flagging, downranking)	Youtube content ID	Twitter quality filter

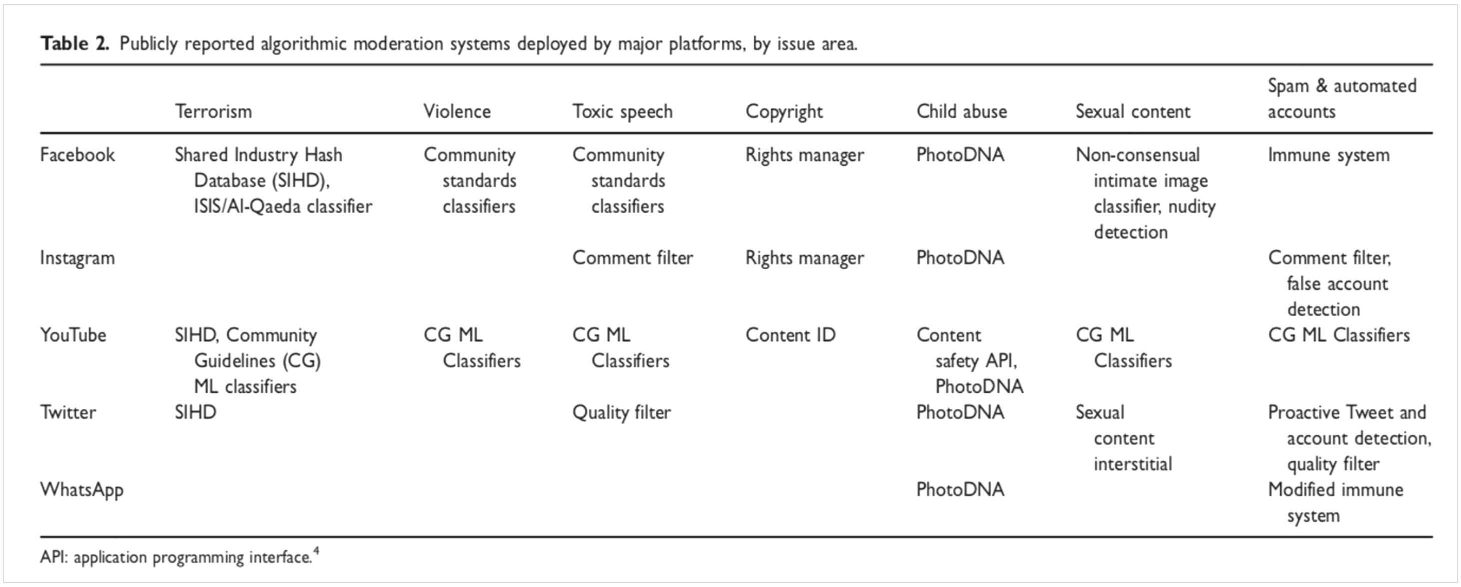


Table 2. Publicly reported algorithmic moderation systems deployed by major platforms, by issue area.

	Terrorism	Violence	Toxic speech	Copyright	Child abuse	Sexual content	Spam & automated accounts
Facebook	Shared Industry Hash Database (SIHD), ISIS/Al-Qaeda classifier	Community standards classifiers	Community standards classifiers	Rights manager	PhotoDNA	Non-consensual intimate image classifier, nudity detection	Immune system
Instagram			Comment filter	Rights manager	PhotoDNA		Comment filter, false account detection
YouTube	SIHD, Community Guidelines (CG) ML classifiers	CG ML Classifiers	CG ML Classifiers	Content ID	Content safety API, PhotoDNA	CG ML Classifiers	CG ML Classifiers
Twitter	SIHD		Quality filter		PhotoDNA	Sexual content interstitial	Proactive Tweet and account detection, quality filter
WhatsApp					PhotoDNA		Modified immune system

API: application programming interface.[4]

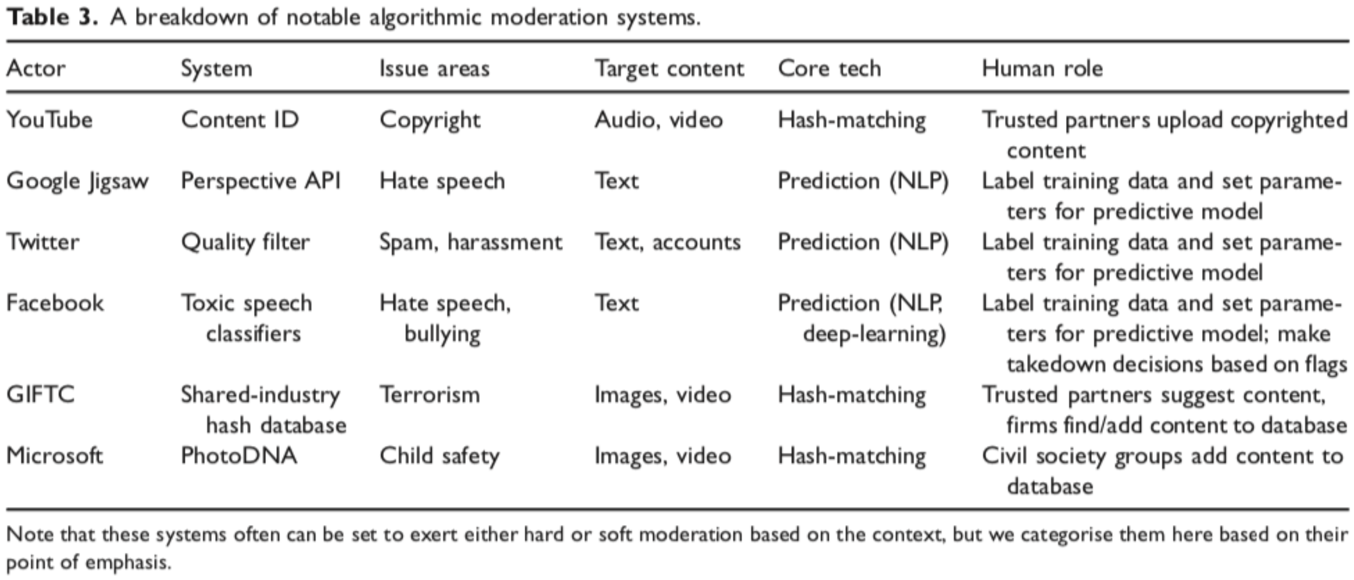


Table 3. A breakdown of notable algorithmic moderation systems.

Actor	System	Issue areas	Target content	Core tech	Human role
YouTube	Content ID	Copyright	Audio, video	Hash-matching	Trusted partners upload copyrighted content
Google Jigsaw	Perspective API	Hate speech	Text	Prediction (NLP)	Label training data and set parameters for predictive model
Twitter	Quality filter	Spam, harassment	Text, accounts	Prediction (NLP)	Label training data and set parameters for predictive model
Facebook	Toxic speech classifiers	Hate speech, bullying	Text	Prediction (NLP, deep-learning)	Label training data and set parameters for predictive model; make takedown decisions based on flags
GIFTC	Shared-industry hash database	Terrorism	Images, video	Hash-matching	Trusted partners suggest content, firms find/add content to database
Microsoft	PhotoDNA	Child safety	Images, video	Hash-matching	Civil society groups add content to database

Note that these systems often can be set to exert either hard or soft moderation based on the context, but we categorise them here based on their point of emphasis.

Like Gillespie, they conclude that even “well-optimized” moderation systems could exacerbate many existing problems by further increasing opacity; further complicating outstanding issues of fairness and justice; and re-obscuring the political nature of speech decisions being executed at scale. **Cobbe (2020)** also offers a substantive evaluation of AI in content moderation, similarly without employing methods or statistical analysis. She similarly argues the emergence of algorithmic censorship is an undesirable development, by permitting platforms to more effectively align public and private online communications while simultaneously undermining those platforms’ function as a space for communication. **Ruckenstein and Turunen (2019)** argue that the current platform logic forces moderators to operate like machines themselves and maintain that the work of CCM should be reframed to integrate moderators’ aims with AI systems; however, they don’t offer much with respect to how that might look in practice, and nothing by way of statistical/data testing. **Einwiller and Kim (2020)** conducted two studies in the U.S., Germany, South Korea, and China to examine how platforms approach moderation—one employing content analysis of policy documents, and another utilizing interviews with representatives of online content providers. They find that content guidelines are often communicated in user-unfriendly terms of service, while only Korean platforms present their policies vividly, and that platforms are not proactive enough in communication for preventing harmful content, often focusing more on avoiding legal ramifications than on educating users. However, Einwiller and Kim (2020) primarily seek to answer what method of *communicating* content moderation policies to users is best, rather than strategies of moderation and their effects

In any case, it is apparent that scholarship on AI and content moderation has only very recently begun to take off, and there is a substantial gap in the literature regarding methodological analyses of the subject.

Flagging

There, of course, is also a user-driven component to content moderation. On platforms, mostly notably Facebook, users can report content violations, specifying how the content violates community guidelines, and why it was reported by the offended. However, how these are processed remains highly opaque, and users cannot know how or on what grounds the platform takes action **(Farkas and Neumayer 2017).** Facebook's report feature mirrors the so-called "flagging" method of content moderation employed by platforms. Flagging describes the mechanisms by which users report offensive content to an online platform, serving as both a solution to the problem of scale and as a rhetorical justification for platform owners removing content. **Crawford and Gillespie (2016)** argue that, though simple in the abstract, in practice flagging is a complex and strategic interaction between users, algorithms, moderators, and platforms themselves. In the aforementioned section on automated social media content analysis tools, **Duarte, Llanso, and Loup (2017)** are adamant that human review of flagged content—whether flagged by users or by automated tools—remains essential for avoiding over-censorship and catching nuances in language use that a classifier might miss.

Unlike the predominately ethnographic and case study-based works on content moderation, **Naab et al (2018)** employed a controlled experiment and a subsequent regression model to test a research question (RQ1)— whether flagging behavior differs when a uncivil comment receive agreement, disagreement, or no response—and three hypotheses—that perceived self-responsibility mediates the effect of agreement or disagreement on flagging behavior (H1); that attribution of responsibility to professional moderators mediates the effect of agreement or disagreement on flagging behavior (H2); and that response politeness moderates the effect of response direction on perceived self-responsibility, attributed professional responsibility, and on flagging behavior (H3). They find that flagging was more likely when intervention information was presented; that providing intervention significantly increases flagging behavior in nearly all groups; and that there exists a significant indirect effect of response direction through attributed professional responsibility on flagging behavior as well as for the interaction of response direction and politeness. Ultimately, they conclude that implementation of a flagging tool not only saves resources but integrates users in the negotiation of shared values.[16]

Not all research on flagging has been positive, though; **Mármol et al (2014)** demonstrate that malicious users may report harmless content just to compromise other users' material and that the large number of users flagging content as offensive makes its management an even more difficult task for commercial content moderation (CCM) workers, in turn, propose a novel reputation-based approach with which social networking sites are capable of automatically assessing accusers' honesty before withdrawing any content.[17]

Labeling

Finally, amidst the toolbox of content moderation options, there is labeling. In order to contextualize content for viewers, platforms may include a visual and/or textual content label to a piece of user-generated content. This practice has become especially prevalent over the past year in response to widespread misinformation regarding both COVID-19 and the U.S. Presidential Election. A cursory search of Instagram will immediately show that any post which even slightly mentions COVID-19 or associated terminologies, regardless of whether it is in fact informational, will include a label with a link reading "Get the latest information from the CDC about COVID-19." **Morrow et al (2020**) term this "contextual labeling," which is not

[16] To answer RQ1, H1, and H2 (Study 1), Naab et al conduct a 2x2x3 between-subjects design, varying intervention information (available vs. unavailable), type of victim (individuals vs. social group), and type of response (agreement vs. disagreement vs. none). They then conduct treatment checks and construct a binary logistic regression model based on the results. To answer H3 (Study 2), they conduct a 2x2 between-subjects design, varying response direction and politeness of the response. To test if the manipulation for response direction was successful, they conduct an analysis of variance (ANOVA), and a second ANOVA with the experimental groups as independent variable to test for the perceived difference in response politeness. They then construct an indirect logistic regression model using PROCESS for SPSS, model 8.

[17] For a comprehensive analysis of the work conducted by

undertaken to correct misinformation, rather, to provide more information to the user that the exact content of a user-generated post does not provide. The alternative to this is "veracity labeling." Twitter, for instance, initially began labeling Trump's post-election tweets with "Other sources have called this election differently," before using more decisive language, and, after January 6th, banning him entirely.

Though labeling has become dramatically more relevant in the past year, such practices nonetheless predate the pandemic and Election. Facebook, for instance, displays a grey screen on many videos, warning viewers of graphic content; in order to see the video, one must click past this cover screen. Anecdotally speaking, such action tends to be taken for videos which are not community standard violations, but nonetheless warrant warning (e.g. videos of George Floyd's murder). As early as 2016, Facebook partnered with third-party fact-checkers at media outlets (Silverman 2016); today, articles which are deemed misinformation by the platform feature a label over the post which reads "False information: Checked by independent fact-checkers" and a button below which provides users the opportunity to "See Why." Psychological scientists **Pennycook et al (2017)** show—based on two controlled studies of around 2,000 MTurkers—that when presented together as in a social media feed, fact-checks of false headlines do reduce belief in misperceptions, but also that untagged headlines (even if false) are perceived to be more accurate, a manifestation of the "implied truth" effect. They find that the magnitude of this effect was roughly 1/3 the magnitude of the "warning effect," whereby misleading headlines with warnings are believed and shared less, and equally large in magnitude as the increase caused by explicitly labeling a headline as true. They find no evidence to support the "backfire effect," whereby substantively correcting false beliefs in the context of news articles may *increase* misconceptions.[18] More recently, **Pennycook et al (2020)** presented evidence in another survey experiment that people share false claims about COVID-19 partly because they fail to think sufficiently about whether or not the content is accurate when deciding what to share; accordingly, they argue platforms should implement some accuracy nudges as they would be

[18] To provide a formal demonstration of why they hypothesize the existence of an implied truth effect, Pennycook et al develop a Bayesian model of belief updating when some headlines are tagged with warnings. They find that a Bayesian will reduce a given person's belief in headlines with warnings and increase belief in headlines without. They provide experimental tests (linear regression, logistic regression) for these model predictions and extend the model to consider the impact of not only adding warnings to headlines found to be false, but also adding verifications to headlines found to be true.

easily scalable (unlike most content moderation efforts) and would not require platforms to make decisions about what content to censor.

Questions persist over the amount of detail that should be in a label. Simple corrections that add a "false tag" contrast starkly with more detailed approaches that offer individuals information in several sentences about why the information is false. **Ecker et al (2020)** tested the effectiveness of simple false tags versus a Twitter-length (140 characters), yet detailed refutation. Though they found that both correction types were effective in correcting misconceptions after one day; however, after a week following exposure and the test, the false tag format was not effective, while the detailed refutation yielded sustained belief change.

Another question scholars (mostly outside of content moderation scholarship) have considered concerning labeling is the degree to which the source of the label and its perceived credibility affects the likelihood misconceptions will be overturned. **Berinsky (2015)** found that a Republican governor or fossil fuel CEO may be better, for instance, at correcting misconceptions about climate change held by Republican-leaning users. **Dias et al (2020)**, meanwhile, found in a survey experiment that emphasizing source information to users had no effect on belief. They followed up this experiment with three more surveys, which suggested a strong, positive correlation between headline plausibility and source trustworthiness. In any case, it is apparent there is substantial disagreement concerning: a) the degree of detail labels should provide in order to rectify misconceptions without causing a backfire effect and b) the degree to which a given label's sourcing determines ultimate belief outcomes.

Bibliography

Anderson, Jessica, Stender, Matthew, Myers West, Sarah, and York, Jullian C. "Unfriending Censorship: Insights from Four Months of Crowdsourced Data on Social Media Censorship." Onlinecensorship.org. March 2016. https://onlinecensorship.org/news-and-analysis/onlinecensorship-org-launches-first-report-download

Artime, Oriol, d'Andrea, Valeria, Gallotti, Riccardo, Sacco, Pier Luigi, and De Domenico, Manilo. "Effectiveness of Dismantling Strategies on Moderated vs. Unmoderated Online Social Platforms." *Scientific Reports*. September 2020. https://www.nature.com/articles/s41598-020-71231-3

Banchik, Anna Veronica. "Disappearing Acts: Content Moderation and Emergent Practices to Preserve At-Risk Human Rights-Related Content." *New Media and Society.* March 2020. https://journals.sagepub.com/doi/abs/10.1177/1461444820912724

Berinsky, Adam J. "Rumors and Health Care Reform: Experiments in Political Misinformation."

British Journal of Political Science. April 2017. https://www.cambridge.org/core/journals/british-journal-of-political-science/article/rumors-and-%20politicalmisinformation/8B88568CD057242D2D97649300215CF2

Binns, Reuben, Veale, Michael, Van Kleek, Max, and Shadbolt, Nigel. "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation." *Social Informatics.* September 2017. https://link.springer.com/chapter/10.1007/978-3-319-67256-4_32

Chancellor, Stevie, Pater, Jessica, Clear, Trustin, Gilbert, Eric, and Choudhury, M.D. "#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities." *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 2016. https://www.semanticscholar.org/paper/%23thyghgapp%3A-Instagram-Content-Moderation-and-in-Chancellor-Pater/d90c289a539842a17c9a19e20142b3b50977273c

Crawford, Kate and Gillespie, Tarleton. "What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media and Society*. 2014. https://journals.sagepub.com/doi/full/10.1177/1461444814543163?casa_token=UN9rDEUVVBEAAAAA%3AScwlAFSdRcWz7sAI4oWj9dF_5Y7viEmGaVc_4lVFt_EFRSmNAPm_HND3QofWuBE0adRORLcf34o

Delort, Jean-Yves, Arunasalam, Bavani, and Paris, Cecile. "Automatic Moderation of Online Discussion Sites." *International Journal of Electronic Commerce*. 2011. https://www.tandfonline.com/doi/abs/10.2753/JEC1086-4415150302

Dias, Nicholas, Pennycook, Gordon, and Rand, David G. "Emphasizing Publishers Does Not Effectively Reduce Susceptibility to Misinformation on Social Media." *Misinformation Review.* 2020. https://misinforeview.hks.harvard.edu/article/emphasizing-publishers-does-not-reduce-misinformation/

Duarte, Natasha, Llanso, Emma, and Loup, Anna. "Mixed Messages? The Limits of Automated Social Media Content Analysis." *Center for Democracy & Technology*. November 2017. https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/

Ecker, Ullrich K.H., O'Reilly, Ziggy, Reid, Jesse S., and Chang, Ee Pin. "The Effectiveness of Short-Format Refutational Fact-Checks." *British Journal of Psychology*. 2020. https://bpspsychub.onlinelibrary.wiley.com/doi/full/10.1111/bjop.12383

Engstrom, Evan and Feamster, Nick. "The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools." *Engine*. March 2017. https://www.engine.is/the-limits-of-filtering

Farkas, Johan and Neumayer, Christina. "'Stop Fake Hate Profiles on Facebook': Challenges for Crowdsourced Activism on Social Media." *First Monday.* September 2017. https://firstmonday.org/ojs/index.php/fm/article/view/8042

Farkas, Johan and Neumayer, Christina. "Disguised Propaganda from Digital to Social Media." *Second International Handbook of Internet Research*. 2020. https://link.springer.com/referenceworkentry/10.1007%2F978-94-024-1555-1_33

Ford, George S. "Notice and Takedown in Everyday Practice: A Review." May 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2963230

Gehl, Robert W., Moyer-Horner, Lucas, and Yeo, Sara K. "Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science." *Television & New Media*. December 2016.

https://journals.sagepub.com/doi/full/10.1177/1527476416680453?casa_token=ryU0oiMnNwwAAAAA%3A3TuX_3QBHy50P3zLLp0hGnIWEVPNbDstYUQMp1g2tVV1Bp_PbFzytP3yZtRMuO1rbKJGuJkpwxY

Gerrard, Ysabel. “Beyond the Hashtag: Circumventing Content Moderation on Social Media.” *New Media & Society*. May 2018. https://journals.sagepub.com/doi/full/10.1177/1461444818776611?casa_token=57nBGNDDb2MAAAAA%3ARjBUu5iWpc8JUAuDg4d6TTmLLF4HTTJfsI8uq0eBZxT89PIBOlqw49ZVpvZ5Oa-BmQJ46DtR1sI

Gillespie, Tarleton. “Platforms Intervene.” *Social Media + Society*. May 2015. https://journals.sagepub.com/doi/full/10.1177/2056305115580479

Gillespie, Tarleton, Aufderheide, Patricia, Carmi, Elinor, Gerrard, Ysabel, Gorwa, Robert, Matamoros-Fernández, Ariadna, Roberts, Sarah T., Sinnreich, Aram, and Myers West, Sarah. “Expanding the Debate About Content Moderation: Scholarly Research Agendas for the Coming Policy Debates.” *Internet Policy Review*. October 2020. https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy

Gorwa, Robert. “What Is Platform Governance?” *Information, Communication & Society* 22. 2019. https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1573914

Gorwa, Robert, Binns, Reuben, and Katzenbach, Christian. “Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance.” *Big Data & Society*. February 2020. https://journals.sagepub.com/doi/full/10.1177/2053951719897945

Gröndahl, Tommi, Pajola, Luca, Juuti, Mika, Conti, Mauro, and Asokan, N. “All You Need is ‘Love’: Evading Hate Speech Detection.” *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. October 2018. https://dl.acm.org/doi/abs/10.1145/3270101.3270103?casa_token=C02P838RnugAAAAA:tSEsEqaQ6mayMsDq8ZEVhyJSV8rZcrhn9XRTRjlk6dVVFGW__hLD3XizDHT_ulJMq-wmy6GLp-Q

Jhaver, Shagun, Appling, Darren Scott, Gilbert, Eric, and Bruckman, Amy. “‘Did you Suspect the Post Would Be Removed?’: Understanding User Reactions to Content Removals on Reddit.” *Proceedings of the ACM on Human-Computer Interaction.* November 2019. https://dl.acm.org/doi/abs/10.1145/3359294

Jhaver, Shagun, Birman, Iris, Gilbert, Eric, and Bruckman, Amy. “Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator.” *ACM Transactions on Computer-Human Interaction* 26, no. 5, article 31. July 2019. https://dl.acm.org/doi/10.1145/3338243

Kaye, David. *Speech Police: The Global Struggle to Govern the Internet*. New York, NY: Columbia Global Reports, 2019. https://globalreports.columbia.edu/books/speech-police/

Keller, Daphne and Leerssen, Paddy. “Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation.” *Social Media and Democracy: The State of the Field , Prospects for Reform* pp. 220-251. Cambridge: Cambridge University Press, 2020. https://books.google.com/books?hl=en&lr=&id=TgH3DwAAQBAJ&oi=fnd&pg=PA220&dq=content+moderation+removal&ots=3BHp00akHt&sig=G6JcJVPOuSw5fEQSTuJ4u0i-9hc#v=onepage&q=content%20moderation%20removal&f=false

Klonick, Kate. “The New Governors: The People, Rules, and Processes Governing Online

Speech." 131 *Harvard Law Review 1598*. March 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937985

Kreiss, Daniel and McGregor, Shannon. "The 'Arbiters of What Our Voters See': Facebook and Google's Struggle with Policy, Process, and Enforcement Around Political Advertising." *Political Communication* 0:1-24. 2019. https://www.tandfonline.com/eprint/9IUFbkGmZ4YHtNXM7sQ2/full?target=10.1080/10584609.2019.1619639

Kreiss, Daniel and McGregor, Shannon. "Technology Firms Shape Political Communication: The Work of Microsoft, Facebook, Twitter, and Google With Campaigns During the 2016 U.S. Presidential Cycle." *Political Communication* 35: 155-177. 2018. https://www.tandfonline.com/doi/full/10.1080/10584609.2017.1364814?casa_token=LVjWiOLdW-AAAAAA%3ADviIRhaGVe_0ZYVYbDG18VlW-1ZuNc6Acx90nR12UgxBkobo_SaG2HuJOQLQMafwq_NWXQiIc-Q

Ledwich, Mark and Zaitsev, Anna. "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization." *Computer Science*. December 2019. https://arxiv.org/abs/1912.11211

Mármol, Félix Gómez, Pérez, Manuel Gil, and Pérez, Gregorio Martínez. "Reporting Offensive Content in Social Networks: Toward a Reputation-Based Assessment Approach." *IEEE Internet Computing*. March-April 2014. https://ieeexplore.ieee.org/abstract/document/6701300

Morrow, Garrett, Swire-Thompson, Briony, Polny, Jessica, Kopec, Matthew, and Wihbey, John. "The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation." *Northeastern University Ethics Institute.* January 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3742120

Munk, Timme Bisgaard. "100,000 False Positives for Every Real Terrorist: Why Anti-Terror Algorithms Don't Work." *First Monday.* September 2017. https://journals.uic.edu/ojs/index.php/fm/article/view/7126

Munn, Luke. "Alt-Right Pipeline: Individual Journeys to Extremism Online." *First Monday*. June 2019. https://firstmonday.org/ojs/index.php/fm/article/view/10108

Naab, Teresa K., Kalch, Anja, and Meitz, Tino GK. "Flagging Uncivil User Comments: Effects of Intervention Information, Type of Victim, and Response Comments on Bystander Behavior." *New Media and Society*. 2018. https://journals.sagepub.com/doi/full/10.1177/1461444816670923?casa_token=aEUQ2IFcGkgAAAAA%3AspTv7sbsezXr5ngyGoKTTfA_pG8Z5mPwmxr1IMje_IRDib1izPqREEraafkLVmTqTiSYjOYSqyM

Naughton, John. "However Extreme Your Views, You're Never Hardcore Enough for YouTube." *The Guardian*. September 23, 2018. https://www.theguardian.com/commentisfree/2018/sep/23/how-youtube-takes-you-to-extremes-when-it-comes-to-major-news-events

Pennycook, Gordon, McPhetres, Jonathan, Zhang, Yunhao, Lu, Jackson G., and Rand, David G. "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention." *Psychological Science*. June 2020. https://journals.sagepub.com/doi/full/10.1177/0956797620939054

Pennycook, Gordon, Bear, Adam, Collins, Evan T., and Rand, David G. "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings." *Management Science*. 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384

Roberts, Sarah T. “Digital Detritus: ‘Error’ and the Logic of Opacity in Social Media Content Moderation.” *First Monday*. March 2018. https://journals.uic.edu/ojs/index.php/fm/article/view/8283

Saka, Erkan. “Big Data and Gender-Biased Algorithms.” *The International Encyclopedia of Gender, Media, and Communication*. July 2020. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119429128.iegmc267

Silverman, Craig. “Facebook Is Turning to Fact-Checkers to Fight Fake News.” *BuzzFeed News*. December 15, 2016. https://www.buzzfeednews.com/article/craigsilverman/facebook-and-fact-checkers-fight-fake-news

Singh, Spandana and Bankston, Kevin. “The Transparency Reporting Toolkit: Content Takedown Reporting.” *New America*. October 2018. https://www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/

Srinivasan, Kumar Bhargav, Danescu-Niculescu-Mizil, Cristian, Lee, Lillian, and Tan, Chenhao. “Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community.” *Proceedings of the ACM on Human-Computer Interaction*. November 2019. https://dl.acm.org/doi/abs/10.1145/3359265

Suzor, Nicolas, Van Geelen, Tess, and Myers West, Sarah. “Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda.” *International Communication Gazette*. February 2018. https://journals.sagepub.com/doi/full/10.1177/1748048518757142?casa_token=ARd6CuwZJJ0AAAAA%3A8T2JCMwySyqNGUt6pbD_Q7R833Z4KPP5la8AawcssdJw8n4qka1LWpV8X9Tayuu1JuCKo53sN4g

Suzor, Nicolas P. *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge: Cambridge University Press, 2019. https://books.google.com/books?hl=en&lr=&id=EjGdDwAAQBAJ&oi=fnd&pg=PR9&dq=suzor++2019&ots=d232EEDMTk&sig=pOZlw1ZUGQJdoJh0VxDDQckkRdQ#v=onepage&q=suzor%20%202019&f=false

Suzor, Nicolas. “Understanding Content Moderation Systems: New Methods to Understand Internet Governance at Scale, Over Time, and Across Platforms.” In Whalen, Ryan (Ed.) *Computational Legal Studies: The Promise and Challenge of Data-Driven Research*. Pp. 166-189. 2020. https://eprints.qut.edu.au/129464/

Tufekci, Zeynep. “Opinion: YouTube, the Great Radicalizer.” *The New York Times*. March 10, 2018. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Urban, Jennifer M., Karaganis, Joe, and Schofield, Brianna. “Notice and Takedown in Everyday Practice.” *UC Berkeley Public Law Research Paper No. 2755628*. March 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755628

7. How Individuals and Groups React to and Evade Content Moderation and How Platforms Respond

The following section provides a more opinionated overview of the literature.
Prepared primarily by LB

Background

Building off of the previous section—which studies strategies for content moderation and their effect on malicious content—the following section overviews scholarship addressing how individuals and groups react to and evade content moderation. Several sub-topics fall within the purview of this research question, such as how users react to content moderation decisions and how that affects their subsequent behavior, how users navigate the appeals process, how individuals and groups circumvent and evade moderation, and the countermeasures platforms take in response. The following section amalgamates and summarizes several of these articles, their methods, and their ultimate conclusion. Once again, section order is not reflective of pertinence, and a bibliography with access links to cited works is included at end of the section.[1]

User Reactions to Moderation Decisions and Subsequent Behavior

While the bulk of scholarship surrounding content moderation decisions concerns the strategies employed by platforms themselves and transparency concerns, some scholars have recently endeavored to explore the other side of the coin—how content moderation decisions affect user experience and subsequent behavior. **West (2018)** conducted a survey experiment of social media users who have experienced a content removal.[2] She found that, though several attributed the reason for the takedown to perceived political bias on the part of the company, the most common theory offered by users was that their content was flagged by another user, usually either someone specific they already knew and/or someone with whom they were in the midst of

[1] There is a substantial volume of work on how individuals and groups evade content moderation in authoritarian states, namely China (**Liao et al 2010; King et al 2013; Hem 2014; Yang 2016; Hobbs and Roberts 2018; Zidan 2018; Cunningham et al 2019; Roberts 2020; Kou et al 2021; Sablosky 2021**). Given the focus of our research thus far, I have not summarized these articles in the present section.

[2] *n*=519; **West** used a page-by-page question design to allow for users to explain in their own words.

a contentious discussion. She further observed that users generally expressed uncertainty as to why the company thought they had violated their policies and how their content was taken down, whether or not they had examined the community guidelines, and lamented the appeals process (elaborated further in the following section). She concludes that the use of automated moderation systems coupled with the overwhelming reliance of platforms on flagging mechanisms to identify content to be removed leaves users confused and frustrated, and reinforces a sense among users that platforms are a place in which they can be targeted for their speech, beliefs, and/or identify.

Jhaver et al (2019a) similarly study user experience with content moderation, asking what kinds of explanations (if any) are provided for removal; whether those explanations matter; how moderated users perceive content removal; and how moderation decisions impact those users' future behavior. The article—which ultimately became a chapter in Jhaver's dissertation **(Jhaver 2020)**—takes an aggregate, quantitative approach as opposed to the survey method employed by West.[3] Employing a massive data set, they identify the frequency at which various removal explanations are given, as well as the frequency at which removal explanations were provided by known bots (~58%, of which ~95% came from "AutoModerator," a moderation tool offered to all subreddits. From this data, they are able to gather 4.7M <user, subreddit> pairs to track future activity. Their regression analyses show that when moderated users are provide explanations, their subsequent post removal rate decreases—whether measured across the entire Reddit platform or within individual subreddits—and their Latent Dirichlet Allocation (LDA) analysis shows that removal explanations are a means through which moderators can begin to develop relationships with moderated users. Based on these findings, they conclude that platforms ought to take a more educational, rather than punitive approach to content moderation to improve community outcomes. When moderator teams commit to transparency and provide

[3] **Jhaver et al** gathered a dataset of ~80M submissions, out of which 17.4M were removed. They adopt the Latent Dirichlet Allocation (LDA) to extract the range of different types of explanations provided through comments. Separately, they employ four logistic regression models to examine the influence of independent variables (past removal rate; explanation rate; average explanation length; explanation through comments rate; explanation by bot rate) upon the two dependent variables (future submission; future removal); they include subreddit subscribers, subreddit submissions, and net subreddit removal rate as control variables.

removal explanations, they argue, users sometimes learn from their mistakes and feel better prepared to make successful future submissions.

Jhaver et al (2019b) employ a more qualitative approach to understanding how content removal affects the individuals who produced it. They conducted a survey of over 900 Redditors—identified via a random sample of submissions—finding that, although 18% accepted that their post removal was appropriate, the vast majority did not—37% did not understand, and 29% expressed some level of frustration about the removal.[4] Both articles (and Jhaver's eventual dissertation) corroborate prior research from the likes of **Kiesler et al (2012)** and **Matias (2019)**—which contends that explicit rules and guidelines increase the ability for community members to know platform norms[5]—and from **Suzor et al (2019)** who argue that individual users whose posts have been removed or accounts suspended need specific explanation about which content breached the rules, how that content was identified, who was responsible for making the decision, and why the conclusion was reached that a rule had been breached.[6] Similar to Jhaver, Suzor et al find that over ¼ of respondents were uncertain about what content triggered a moderation decision, and even more had insufficient information to understand why a

[4] **Jhaver et al** gathered a random sample of 10K submissions of Reddit using PRAW Reddit API. Since this doesn't allow for direct retrieval of random submissions, they start with randomly sampling a subreddit and retrieving the most recent submission. Then, they sent custom surveys to those who were moderated via private message. They conduct regression analyses on their DVs, 'fairness' and 'post again', adding control variables and posting context variables. They find that reading the rules was significantly associated with fairness even after controlling for demographics and prior Reddit history variables, but in the opposite direction expected; when users read the rules, they are less likely to consider the removal fair. They also find that when users perceive the rules to be clear they are more likely to see the removal as fair and more likely to post again.

[5] **Matias** randomized announcements of community rules to large-scale conversations in a science-discussion community (r/science) with 13M subscribers to test whether displaying community rules could reduce concerns about harassment, finding that displaying the rules increased newcomer rule compliance by >8% and increased the participate rate of newcomers in discussions by 70% on average.

[6] **Suzor et al** conducted a survey analysis of 380 responses submitted by users who had been adversely affected by the removal of content they posted on social media platforms or by the suspension of their account. They analyze complaints and apparent confusion evident within the responses to identify deficiencies in the way platforms communicate with users about moderation decisions.

decision was made; only half expressed confidence in their understanding of the platform's moderation action. Relatedly (albeit somewhat tangentially), **Gibson (2019)** analyzed ~250K comments on Reddit from two forums over a 14-month period with differing moderation policies ("safe space" and "free space"), finding that the former spurred more moderator action and self-censorship, fostering more positive discussions.[7]

Chandrasekharan et al (2017) also looked at content moderation on Redditors behavior, studying Reddit's widely-publicized 2015 ban of r/fatpeoplehate and r/CoonTown [sic] in terms of its effect on both participating users and affected subreddits. Using a dataset of over 100M Reddit posts and comments, they generate hate speech lexicons to examine variations in hate speech usage, finding that the ban ultimately worked for Reddit.[8] They also found that more accounts discontinued using the site than anticipated, and that among those who stayed, their hate speech usage decreased by at least 80%; the subreddits which saw an influx of "migrants" from the two subreddits saw no significant changes in hate speech usage. Accordingly, they argue that—since Reddit's decision to ban the two subreddits dispersed participants to other parts of the site and reduced overall hate speech usage—platforms should ban the spaces where deviant groups congregate. **Srinivasan et al (2019)** also uncovered positive findings with respect to user responses to moderation; by leveraging a new quasi-experimental design they call "delayed feedback" upon a case study of the "ChangeMyView" subreddit, they show that individuals subject to moderators (comment removal) in turn regulate their bad behavior, reduce

[7] **Gibson** compiled a data set of ~300K comments made on two subreddits and analyzed their language using Linguistic Inquiry and Word Count (LIWC2015), which sorts and displays the words present in text into a variety of categories (e.g. linguistic, social/affect, and cognitive) by frequency. The posts are then filtered at the author and comment levels for crossposters. They then conduct simple linear regressions to model the rate at which each subreddit gained subscribers and constructed linear mixed models fit by REML for each linguistic dimension to minimize error.

[8] **Chandrasekharan et al** collected all posts made in 2015 for the two subreddits using the Reddit 2015 Corpus. They employ Automatic Keyword Identification: SAGE Analysis to identify hate speech. They employ causal inference techniques such as matching the treatment subreddits to control subreddits that could have been banned, matching the treatment subreddit users to control subreddit users with similar posting behavior, using difference-in-difference procedure to compare the pre- and post- differences between the treatment and control groups, and an interrupted time series analysis.

toxicity in language, and reduce noncompliance, identifying a clear causal effect in the latter case.[9]

Most recently (to be published October 2021), **Jhaver et al (2021)** evaluate the effectiveness of deplatforming as a moderation strategy on Twitter. By conjoining a case study of high-profile influencers deplatformed on Twitter—Alex Jones, Milo Yiannopoulos, and Owen Benjamin—with a data set of over 49M tweets, they demonstrate that deplatforming significantly reduced the number of conversations about all three individuals on Twitter. They further found, by analyzing Twitter-wide activity of those influencers' supporters that the overall activity and toxicity levels of those supporters declined after the deplatforming. They also contribute the following methodological framework for systematic examination of the effectiveness of moderation interventions and the effects of deplatforming, reproduced below:

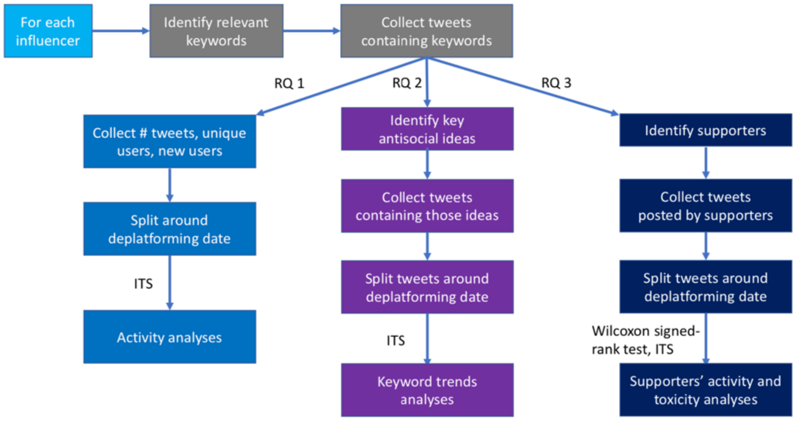

For each influencer
Identify relevant keywords
Collect tweets containing keywords
RQ 1
RQ 2
RQ 3
Collect # tweets, unique users, new users
Split around deplatforming date
ITS
Activity analyses
Identify key antisocial ideas
Collect tweets containing those ideas
Split tweets around deplatforming date
ITS
Keyword trends analyses
Identify supporters
Collect tweets posted by supporters
Split tweets around deplatforming date
Wilcoxon signed-rank test, ITS
Supporters' activity and toxicity analyses

Fig. 1. Methodological framework used in this study to examine how deplatforming affects (1) change in activity around influencers (RQ 1), (2) change in the spread of ideas associated with influencers (RQ 2), and (3) change in overall activity and toxicity levels of their supporters on the platform (RQ 3).

[9] To identify the causal effects of moderation in the case study and account for the inability of standard interrupted time-series (ITS) analysis to prove that a user, after making a problematic comment, would have behaved worse if the comment had not been removed, Srinivasan et al leverage their "delayed feedback" approach. This exploits the observation that some affected individuals make comments in the (often hours-long) "pre-removal" window wherein the to-be-removed comment has not yet actually been removed.

The Appeals Process

In addition to studying how moderation decisions affect subsequent user behavior, some scholars have looked at how aggrieved users go about seeking redress for perceived unwarranted removal. In her aforementioned survey, **West (2018)** found that roughly half of users appealed the decision to remove their content; of the half that did not, a majority claimed they either did not know how to go about the appeal or did not expect any response. She further observed that users expressed frustration that they were not offered an opportunity for human interaction in seeking redress for the removal of their content, a frustration which was reinforced when the appeals process did not result in a resolution of the problem. In many cases, the content would go through the same review process, not incurring any additional scrutiny of the content.

Vaccaro et al (2020) ask specifically how users experience contesting unfavorable social media content moderation decisions and provide some contextual background to the appeals processes operated by major platforms.[10] For instance, when users appeal an account suspension, Facebook asks them to provide minimal information (email address, full name, and an image of an ID) to confirm their identity; the "appeal" is the action of sending the request, and the system does not provide a space for users to explain. Employing a between-subjects design to study the impact of different strategies/processes for appeal, they unexpectedly find that none of the appeal designs proposed to improve users' perceptions of the Fairness, Accountability, and Trustworthiness and feelings of Control (FACT) compared to the "no appeal" baseline. They nonetheless suggest based on qualitative findings that platforms should modify appeals systems to improve user experience.

Lyons, Velloso, and Miller (2021a; 2021b; 2021c) have written several qualitative articles recently about contestability as an ethical principle as it relates to algorithmic moderation decisions. Based on an analysis of submissions made in response to a discussion paper released by the Australian government, they find that there is no clear consensus on the true nature of contestability; for some, it is an ethical principle in its own right and for others it's a means to an

[10] **Vaccaro et al** use Qualtrics to identify participants, and conduct a between-subjects design involving a scenario where participants' Facebook account has been suspended, with experimental conditions varying based on whether and how they can appeal. The experimental intervention is introduced via a mock-up of an interface, which has three sections. The types of appeal tested are: *none, written/human, written/algorithm, and behavioral/algorithm*.

end. They argue that to reduce negative consequences, AI systems must be responsibly designed, developed, and deployed; though the ability to contest decisions is not the only mechanism to ensure fairness in AI systems, it is a crucial safeguard.

Relatedly, **West (2017)** examined the strategies and tactics of users turning to collective action—often in part due to the inadequacies of the appeal process—against platforms for moderation policies, using the "#FreetheNipple" movement as a case study. The movement, which ballooned in the early 2010s, led to quiet changes in Facebook's community guidelines to allow exceptions for breastfeeding mothers and mastectomy scarring. Some protesting the continued policy (despite the narrow changes) have subverted it by applying a "digital pasty," a kind of moderation evasion akin to those unpacked in the following section.

How Individuals and Groups Evade Detection

Content moderation policies and methods are in a state of constant evolution, as individuals and groups are regularly identifying new ways to evade detection by platforms. Groups leveraging online content to commit offline harms are often early adopters of tactics to circumvent technologically-oriented limitations. For example, ISIS fighters created the "Syrian Archive," a software that would download new material from Syrian YouTube channels to its servers before they were purged from the site, forcing the platform to ramp up moderation to an extent at which it could delete videos at a speed faster than the Archive could download them. The YouTube "scraper" was subsequently implemented in December 2016; every ten seconds, the scraper builds a sample of 50 videos which were published in a ten second window one-minute prior. In a study of the scraper and similar, large-scale content moderation systems, **Suzor (2020)** found that—assuming removed videos are close to average duration—the sample of videos tested is likely close to 10% of total videos uploaded to the platform, and of those, approximately 6.7% have since been removed because of some form of content moderation.[11] In

[11] Suzor collects random samples of pubic social media content, and then tests the availability of that sample at a later date; he attempts to assemble the initial sample as close as possible to the time a post is uploaded or made public, test availability again between 8-20 days later, and then re-test it two months after it was initially collected (to identify posts that were reinstated). However, the article only reports on the content that was a) captured within an hour of being posted and b) was tested again between 8-20 days later.

the same article, Suzor (2020) also looks at Twitter's tester, which has been running stably since January 2018. The Twitter tester follows the same logic as the YouTube tester, collecting a random sample of tweets as close as possible to the time they were posted. He argues that the error results that Twitter returns when it moderates content are, compared to YouTube, very basic; whereas YouTube data allow for categorical analysis of alleged malfeasance, when Twitter removes a tweet it typically applies a sanction to the user's account which will only be lifted after the offending tweet is deleted or appeal is won. Suzor's article also studies Instagram's most visible and controversial mechanism to moderate content, as well as a more traditionally computer science-oriented article with Gray **(Gray and Suzor 2020)**, which employs a dataset of ~77M YouTube videos to explore how digital and computational methods can be improved and leveraged in large-scale moderation systems.[12]

Extremist groups—particularly jihadists—evolve continuously in such manners to circumvent monitoring and enable content to emerge despite the increasing implementation of artificial intelligence (AI). For more on terrorist organizations—particularly outside of content moderation scholarship—see the final section, *Additional Extremism Concerns and Platform Countermeasures.*

The challenge platforms face in identifying methods and practices of moderation circumvention is certainly not confined to anti-extremism and counter-terrorism. Given that so much of the labor behind content moderation is now automated, a prominent method of evasion is lexical variation. A substantial body of literature has been devoted to studying moderation circumvention using the case study of pro-eating-disorder communities, which platforms such as Instagram and Pinterest took an uncharacteristically vocal stand against. **Chancellor et al (2016)** tested the effectiveness of Instagram's hashtag bans on removing pro-ED content (see *Hashtag Bans* in previous section), finding that the pro-ED community adopted non-standard lexical variation of moderated tags to circumvent these restrictions, and these variations grew increasingly complex over time (e.g. "#thyghgapp). Ultimately, they argue that despite

[12] Gray and Suzor use the BERT language model to train a machine learning classifier to identify videos in categories that reflect ongoing controversies in copyright takedowns. They find that a simple neural network infrastructure, a pre-trained BERT language model, and substantial cloud processing power are able to achieve satisfactory performance on a multiclass classifier over short texts with a relatively small number of training examples.

Instagram's moderation strategies, pro-ED communities are active and thriving, demonstrating its insufficient effectiveness as an intervention for communities of deviant behavior.[13] **Gerrard (2018)** studied the same community—extending to sites like Pinterest and Tumblr and adopting a more qualitative approach—finding not only that untagged pro-ED content can easily be found and that users are evading hashtags, but also that platforms' algorithmic recommendation systems in fact recirculate pro-ED content, revealing the limitations of hashtag logics in social media content moderation.

Scholars using different case studies have similarly found that individuals and/or groups purveying undesirable content act dynamically in their efforts to evade moderation. **Ferrerira (2021)** found that during Brazil's 2018 elections—based on a content analysis of the 153 false or misleading narratives most shared—that disinformation campaigns were able to adapt to new platform control measures as they emerged and meet the most recent audience preference.

When **Gröndahl et al (2018)** reproduced seven state-of-the-art hate speech detection models from prior work, they found that all proposed detection techniques are brittle against adversaries who can automatically insert typos, change word boundaries, or add innocuous words to the original hate speech. The categorize common methods of "attack" as follows: **word changes** (e.g. inserting typos, writing words with numbers); **word-boundary changes** (e.g. inserting whitespace, removing whitespace); and **word appending** (e.g. appending common and non-hateful words). Similarly, **Zhu et al (2021)** explain that fringe groups and organizations have a long history of using **euphemisms** to conceal what they are discussing, and that existing tools for enforcing policy automatically—which rely on keyword searches for words on a "ban list"—are notoriously imprecise. This is additionally challenging for moderators because, as with lexical variation, the group using a given euphemism need only invent another one, keeping moderators one step behind. Another tactic employed by malicious video uploaders which **Singh et al (2019)** examined is to place violent or sexually explicit scenes within a video, often to catch

[13] Chancellor et al analyze lexical, behavior, and topical changes associated with the emergence of lexical variation in pro-ED communities. They offer a set of general rules to define lexical variants, and categorize them as root tags and tag chains, the set of all lexical variants of a given tag. They used Instagram's official API to collect >8M public posts in the pro-ED space; to account for posts including banned tags, which are not returned in the API, they sample for pro-ED tags that co-occurred with banned tags and create a candidate pro-ED post set.

children's' attention. These "child unsafe" scenes are usually confined to only a few frames per video, making automated detection difficult.[14]

Migration to Other Platforms/The "Dark Web"

Apart from the tactics employed by individuals and groups to evade moderation *within* certain communities (e.g. word changes), some scholars have studied the migration of those producing unwanted content (particularly extremists) to other platforms and and the unique challenges this trend fosters. As **Heller (2019)** elaborates, scrutiny from tech giants has pushed many online extremists to the "Dark Web," or to fringe platforms like 4chan, 8kun (formerly 8chan), and Gab as communications channels, forums which have few if any restrictions on hate speech, disinformation, and other types of malicious content. They have often coupled this migration with "outlinking," spreading content on platforms using links to other platforms where the malicious content can be found (e.g. justpaste.it, sendvid.com, and archive.org).

Gab—a new social network marketed as an alternative to Twitter—in particular plays a pivotal role in this conversation. After the major platforms de-platformed white supremacist individuals and groups, the alt-right movement needed an accompanying "alt-tech" movement to help fill the void and maintain an internet infrastructure. Of the smaller platforms competing for attention in the fall of 2017 in the wake of the Unite the Right Rally in Charlottesville, Gab emerged as a publicly vocal supporter of free speech and as an expressly unmoderated online space, positioning itself to take on users abandoning Twitter in three overlapping movements: the free speech movement, the open technology movement, and the Alt-Right (**Donovan, Lewis, and Friedberg 2019**). Absent regulatory intervention, platforms such as Google Play Store and the Apple App Store have taken it upon themselves to remove (or, in Apple's case, reject) the social network from its systems. According to Gab guidelines, the site does not allow illegal pornography nor the promotion of violence or terrorism; everything else is fair game.

Jasser et al (2021) recently published an empirical analysis of Gab, finding that the technological affordances it boasts—a lack of content moderation, a culture of anonymity, and

[14] **Singh et al** leverage Long Short Term Sectionry (LSTM) based autoencoder to learn effective video representations of video descriptors obtained from VGG16 Convolutional Neural Network (CNN). They then create a dataset of ~110K video clips curated for child unsafe content, finding that the deep learning approach is much more effective.

its microblogging architecture and funding model—have indeed fostered an ideologically far-right and often hateful community, albeit one dwarfed by major platforms and among which explicit calls for violence are rarer than one might expect.[15] **Zhou et al (2019)** further find that discourse on Gab may be driven by a small number of superparticipants; that users do not seem to take advantage of its permissive environment to engage in deliberative discussion, but instead seem to broadcast more and engage less; and that there are a significant connection between Gab and German language speakers, and there is a significant degree of content on the platform originating from Russian state-sponsored sources.[16] **Lima et al (2018)** conducted a similar analysis, finding that Gab hosts known banned users from other social networks; that the majority of users are conservative, white men; and that most news domains shared within Gab are not popular in other social media or the Internet as a whole.[17] **Zannettou et al (2018)** conducted the earliest of the now numerous scholarly investigations of Gab, finding that the platform is predominately used for the dissemination and discussion of news and world events; that it attracts alt-right users, conspiracy theorists, and trolls; and that the prevalence of hate speech on the platform is much greater than Twitter, but lower than 4chan's "Politically Incorrect" board.[18]

Another platform worth mentioning is Parler, which similarly promotes its service as allowing for free expression without fear of deplatforming and became quickly popular among

[15] **Jasser et al** collect data from Gab's API using a snowball methodology. They first obtained data for the most popular users, then iteratively collected data from their followers and the accounts they were following. This ultimately led to a dataset of over 47M posts. They also constructed separate samples of posts by far-right platform "elites." They then analyzed the posts using qualitative content analysis.

[16] **Zhou et al** utilize Gab's sequential message ID system to obtain every message (~17M) during the collection period and lists of accounts followed-by and following for each user (~146K). They then conduct a comparative network analysis between Gab and Twitter.

[17] **Lima et al** crawled users and their posts by following a Breadth-First Search (BFS) scheme on the graph of followers and friends, using as seeds users who authored posts listed by categories in the Gab main page. They also collect URLs' headlines of posts categorized as news.

[18] **Zannettou et al** use Gab's API to crawl the social network using snowball methodology and obtain data for the most popular users and iteratively collect data from all their followers and their followings. They then calculate summary statistics on their findings as well as data visualizations, such as a CDF of followers and following and temporal analysis of Gab posts.

those who were suspended from mainstream social networks, particularly following platforms' efforts to curb misinformation surrounding the 2020 Election and the storming of the U.S. Capitol on January 6, 2021. Parler grew rapidly in large part because the service was endorsed by several conservative public figures. It has since been progressively deplatformed, as its app was removed from Apple and Google Pay stores and the website was taken down by the hosting provider. **Aliapoulios et al (2021)** provide an early analysis of the network, overviewing its function and presenting a basic characterization of a large data set of posts. They find, as expected, that the platform witnesses large influxes of new users after being endorsed by popular figures and in reaction to the 2020 Presidential Election, and further show that discussion on Parlor is dominated by conservative topics, President Trump, and at times conspiracy theories such as QAnon.[19]

Another platform which—despite modest efforts to combat hate speech—to which extremists have migrated is Discord, a real-time, voice-based communication intended for gaming [see **Glaser (2018)** for *Slate*]. **Jiang et al (2019)** investigate how such new modes of interacting can introduce such unexpected challenges—not only are there new ways to break rules that moderators of text-based communities find unfamiliar, such as disruptive noise and voice raiding, but acquiring evidence of rule-breaking behaviors is much more difficult.

Jardine (2019) outlines what he calls the "surface-to-Dark Web content cycle," a four step process in which content is produced on the "surface web," is then restricted on the surface web, is displaced to the Dark Web, from where it percolates back to the surface web. He also outlines two easier pathways by which malicious content can resurface: 1) through direct resurgence on the surface web (e.g. *The Daily Stormer* trying to find a foreign domain to host its content after being delisted) and 2) by way of multi-media content pieces (e.g. memes, infographics, and slogans). The frequency at which this occurs is unknown—particularly percolation through idea networks; however, Jardine demonstrates that malicious content can make its way back onto platforms no matter how aggressively said platform retaliates against the offensive post and/or offender.

[19] **Aliapoulios et al** use a custom-built crawler to access the undocumented by open Parler API. They collect all user profile information for the 13.25M Parler accounts created between August 2018 and January 2021, as well as 98.5M public posts and 84.5M public comments from a random set of 4M users. They find that 2.09% of users are banned.

Additional Extremism Concerns and Platform Countermeasures

A significant portion of relevant research on the online activities of extremist/terrorist organization falls outside the bubble of online content moderation scholarship, often into fields which rarely, if ever, overlap with content moderation apart from this issue, such as terrorism studies, defense studies, etc. The first jihadist websites surfaced around 2000, and by the end of the decade web forums became the primary location for online meetings and jihadist hubs. As early as 2008, some jihadists advocated leveraging mainstream platforms such as Facebook and YouTube; in the wake of the Paris terror attacks, social media platforms began to intervene more. According to **Bindner and Gluck (2017)**, ISIS subsequently migrated towards Telegram, an encrypted messaging application, in essence choosing operational security and longevity over its communication reach and visibility.

Within social media, jihadists employing a myriad of ever-evolving tactics to circumvent monitoring, many of which align with what Gröndahl et al find in their aforementioned computer science article; this includes **image distortion or cropping (logos)** in order to cheat AI, **hijacking hashtags**, posting egregious content in **comments instead of main posts** to avoid detection, using **profile pictures, bios, and names** not openly suggesting a link to the jihadist movement**, "hiding" media** in multiple uploads in a single post, use of **disappearing "stories" (Bindner and Gluck 2020)**. Though Telegram as a platform has attempted to cub IS content, **Clifford and Powell (2019)** (of GWU's program on extremism) found that the platform is still the group's preferred platform of jihadists to date.

However, another tactic jihadists have adopted to obfuscate moderation is to diversify their media outlets on the surface web. Recently, IS content has emerged on Yahoo Together, Viber, TamTam, Baaz, and Zellos, and on various other decentralized platforms **(Bodó 2018)**. Bindner and Gluck argue this fragmentation and the increased monitoring on mainstream social media have curbed the viral effect of their malicious material and isolated it from the public to a certain extent, but with the consequence that this pattern makes it less traceable by authorities, who end up playing cat and mouse but running in all directions.[20]

[20] For a strong, wide-reaching, qualitative look at extremist (not exclusively jihadist) content, I highly recommend *Extreme Digital Speech: Contexts, Responses, and Solutions*, ed. Bharath Ganesh and Jonathan Bright, from *Vox-Pol*, within which the article cited from **Bindner and Gluck (2020)** is included.

But just as extremist groups purveying malicious content have continually evolved in their capacity to obfuscate moderation, platforms have continually evolved in their capacity to detect and remove unwanted material. As with any subtopic under the umbrella of content moderation, advancements in AI and automated detection technology are pertinent to counterterrorism on the internet. The primary automated tools used to curate, organize, filter, and classify information in the online content moderation process are as follows **(Singh 2019)**:

i) **Digital Hash Technology**—which works by converting images and videos from an existing database into grayscale format, overlaying them onto a grid, and assigning each square a numerical value. The designation of a numerical value converts the square into a hash, or digital signature, which remains tied to the image or video and can be used to identify other iterations of the content either during ex-ante moderation or ex-post proactive moderation. Digital hash technology has been widely adopted by internet platforms. Child Sexual Abuse Material (CSAM) detection technology, known as PhotoDNA, a hash technology, has expanded to become a powerful tool, even creating newer technologies like ContentID, a YouTube technology which allows users to create digital hashes for the content to protect against copyright violations. Despite its power and apparent effectiveness, several concerns have arisen (elaborated upon further in this section), such as the definition of extremist content, bias in algorithmic training (e.g. reduced reliability of automated tools trained to focus on IS or al-Qaeda in addressing the larger corpus of extremist content), and the lack of transparency and accountability around how digital hash technology is deployed.

ii) **Image Recognition**—which is employed by digital hash technologies, but can also be used more broadly in the moderation process. During ex-post proactive moderation, for instance, image recognition tools can identify specific objects within an image (e.g. weapons) and decide bsed on factors including user experience and risk whether the image should be flagged to a human for review. Such automated recognition tools are currently employed by platforms to filter through and prioritize cases for CCM workers. Concerns exist surrounding such technologies' frequent inability to incorporate more nuanced and contextual insights into their detection procedures, the quality of datasets models are trained on, and the lack of transparency around how databases are compiled, what types of content they focus on, how accurate they are across different categories of content, and how much user expression has been accurately and erroneously removed as a result.

iii) **Metadata Filtering**—which utilizes files' metadata, the descriptive characteristics about content. Such filtering tools can be used during ex-ante and ex-post proactive moderation to search a series of filed to identify content that fits a particular set of metadata parameters. This tool is used particularly to identify copyright-infringing materials. However, because metadata manipulation and mislabeling can be easy, the effectiveness and accuracy of such tools are limited, particularly compared to the others.

iv) **Natural Language Processing (NLP)**—a set of techniques that uses computers to parse text; in the context of content moderation, text is typically parsed in order to make predictions about the meaning of the text, such as what sentiments it indicates. NLP classifiers are particularly used to detect hate speech and extremist content, and to perform sentiment analysis. NLP classifiers are generally trained on text examples that have been annotated by humans in order to indicate whether they belong to a particular category (e.g. extremist v. non-extremist). Though platforms have been increasingly exploring and adopting NLP classifiers, the technology is still limited for several reasons: a) NLP technologies are domain-specific, which means they can only focus on one particular type of objectionable content; b) because there is significant variation in how speech is expressed, these categories are narrow; c) finding and compiling comprehensive enough datasets to train NLP classifiers is challenging, expensive, and tedious; d) in order for NLP classifiers to operate accurately, they need to be provided with clear and consistent parameters and definitions of speech.[21]

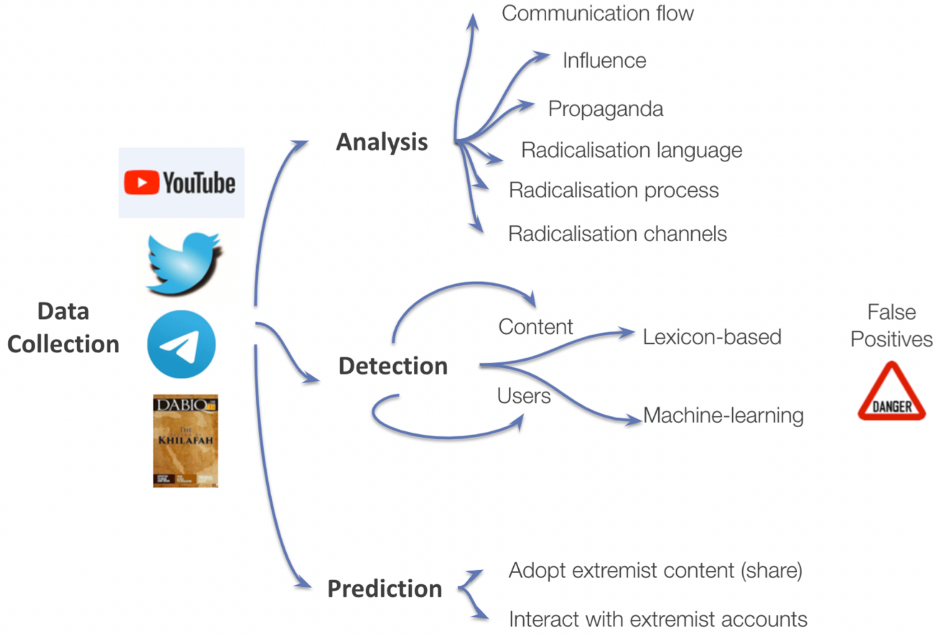
Llansó et al (2020) describe automation in content moderation succinctly, as such: it can be used in the related phases of proactive *detection* of potentially problematic content and automated *evaluation* and *enforcement* of a decision to remove, tag/label, demonetize, demote, or prioritize content, and can separately be employed to automate the **g***eneration* of content and accounts (e.g. bots). In an article published just this week, **Borelli (2021)** argues that—since the emergence of IS—the firms she studied (Facebook, Google, YouTube, and Twitter) have displayed agency and creativity in their approximation of this new responsibility, going above and beyond what is legally required and expected of them. Their self-regulated commitment to counterterrorism and countering violent extremisms (CT/CVEs), she maintains, effectively makes them actors of counterterrorism.

In order to combat migration of extremists from one platform to another, technology firms (initially Google, Microsoft, Facebook, and Twitter) set up the Global Internet Forum to Counter Terrorism (GIFCT) to create a shared database of extremist content to block repeated uploads of the same material. They do so by creating a cryptographic hash (i.e. a digital fingerprint) of each identified image or video clip, automatically blocking database matches that

[21] I highly recommend exploring this thorough report from **Singh (2019)**, "Everything in Moderation: An Analysis of How Internet Platforms are Using Artificial Intelligence to Moderate User-Generated Content."

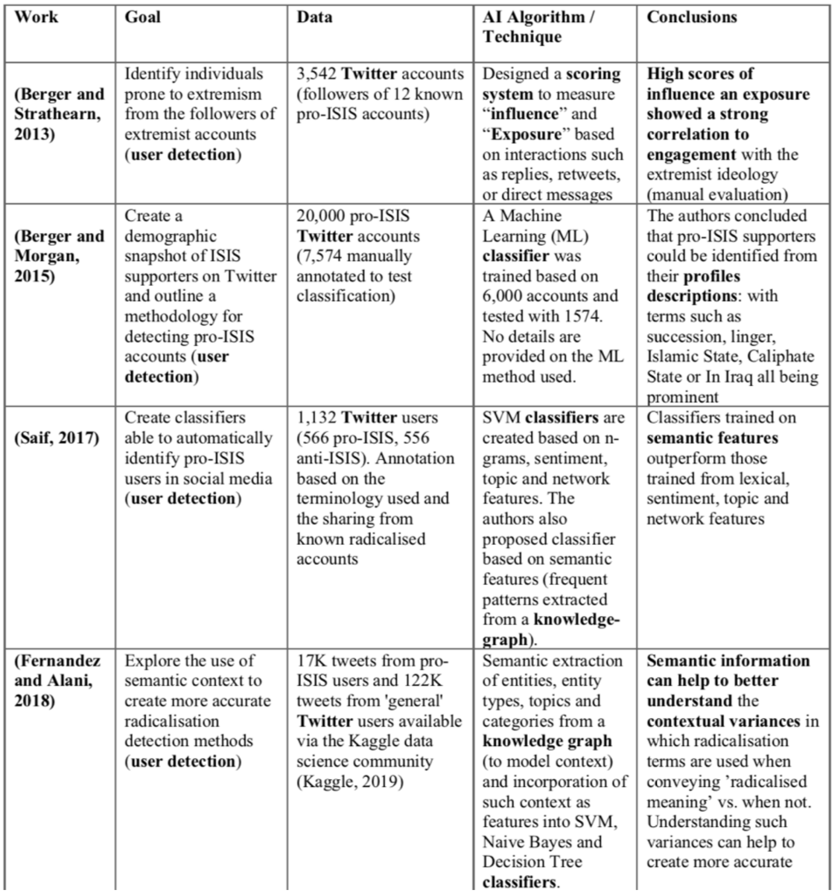
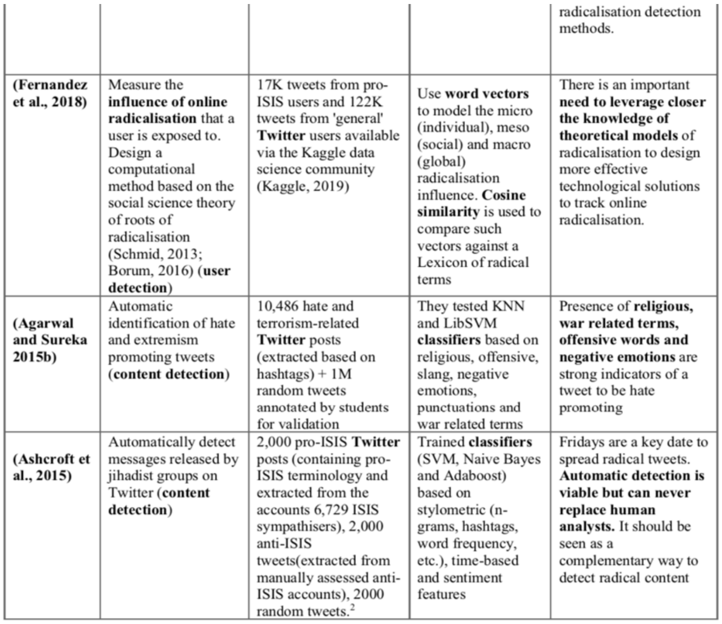
are found when content is uploaded to any of those platforms. As of 2019, the database contained 88K digital fingerprints, and smaller platforms like Instagram, LinkedIn, and Snapchat have been added. In a report for the Transatlantic Working Group, **Heller (2019)** concedes that early data indicate that the GIFCT's hash-sharing efforts are working, at least based on what information we do have, and if content removal is the metric of success. Statistics from the GIFCT seem to reinforce its claim that it is effective at removing a small but significant slice of content which accounts for the worst types of terror-related content. Following a series of interviews with GIFCT partner companies and law enforcement, **van der Vegt et al (2019)** similarly argue that this database of hashes—focusing on the most extreme and egregious terrorist images and videos—should be expanded to include more borderline content.

The UK government, in partnership with the AI firm ASI Data Science, announced a similar approach in 2018, employing machine learning to analyze a range of subtle patterns within videos and determine whether they match patterns found in IS propaganda. Able to distinguish original content from videos that discuss the same imagery in different contexts such as news reporting, this tool is reportedly 94% accurate, with a false positive rate of 0.005%. Such a tool attests to the power AI offers in the realm of online counterterrorism, as such a tool would take approximately 20,000 people to perform the same tasks; nonetheless, the program does not remove content on its own, but does flag videos for human consideration **(Gallacher 2019).** Platforms themselves have found automated detection to be immensely valuable as a tool; speaking on behalf of Facebook, **Bickert and Fishman (2018)** note that 99% of IS content was removed before reported to the platform by a user. Similarly, YouTube reported on their official blog that 98% of the videos removed for violating violent extremism rules are detecting by an automated system. In a recent article, **Fernandez and Alani (2021)** produce the following flow chart overviewing AI approaches to counter online radicalization:



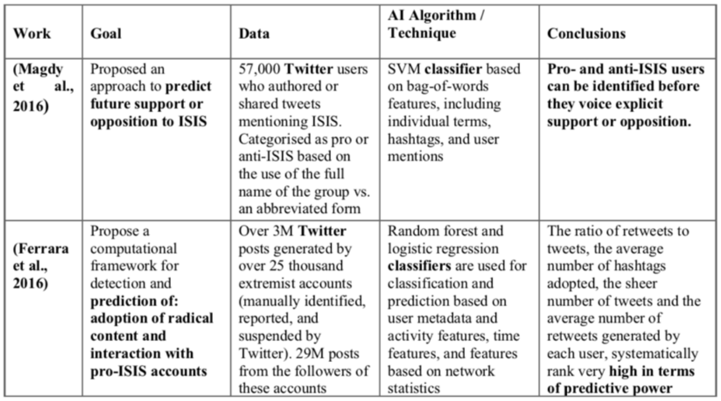
Nonetheless, as Gallacher notes**, challenges for platforms** still exist: a) how to deal with false positives; b) how to prevent malicious attacks on the tools themselves; c) how to apply these tools when groups move to more encrypted spaces and smaller platforms; and d) how to appropriately apply tools built with a focus on one domain of extremism to other forms. **Fernandez and Alani (2021)** frame the challenges facing AI in content moderation a bit differently, as follows: a) the lack of a common definition of prohibited and extremist internet activity; b) the lack of solid verification of the datasets collected to develop detection and prediction models; c) the lack of cooperation across research fields, since most of the developed technological solutions are neither based on, nor do they take advantage of existing social theories and studies of radicalization; d) the constant evolution of behaviors associated with online extremism in order to avoid being detected by the developed algorithms; and e) the development of ethical guidelines and legislation to regulate the design and development of AI technology to counter radicalization. They also offer useful tables on scholarly work which has thus far focused on AI as it pertains to online radicalization, copied below:

Table 2: Approaches that focus on the **detection** of online radicalisation

Work	**Goal**	**Data**	**AI Algorithm / Technique**	**Conclusions**
(Berger and Strathearn, 2013)	Identify individuals prone to extremism from the followers of extremist accounts (**user detection**)	3,542 **Twitter** accounts (followers of 12 known pro-ISIS accounts)	Designed a **scoring system** to measure "**influence**" and "**Exposure**" based on interactions such as replies, retweets, or direct messages	**High scores of influence an exposure showed a strong correlation to engagement** with the extremist ideology (manual evaluation)
(Berger and Morgan, 2015)	Create a demographic snapshot of ISIS supporters on Twitter and outline a methodology for detecting pro-ISIS accounts (**user detection**)	20,000 pro-ISIS **Twitter** accounts (7,574 manually annotated to test classification)	A Machine Learning (ML) **classifier** was trained based on 6,000 accounts and tested with 1574. No details are provided on the ML method used.	The authors concluded that pro-ISIS supporters could be identified from their **profiles descriptions**: with terms such as succession, linger, Islamic State, Caliphate State or In Iraq all being prominent
(Saif, 2017)	Create classifiers able to automatically identify pro-ISIS users in social media (**user detection**)	1,132 **Twitter** users (566 pro-ISIS, 556 anti-ISIS). Annotation based on the terminology used and the sharing from known radicalised accounts	SVM **classifiers** are created based on n-grams, sentiment, topic and network features. The authors also proposed classifier based on semantic features (frequent patterns extracted from a **knowledge-graph**).	Classifiers trained on **semantic features** outperform those trained from lexical, sentiment, topic and network features
(Fernandez and Alani, 2018)	Explore the use of semantic context to create more accurate radicalisation detection methods (**user detection**)	17K tweets from pro-ISIS users and 122K tweets from 'general' **Twitter** users available via the Kaggle data science community (Kaggle, 2019)	Semantic extraction of entities, entity types, topics and categories from a **knowledge graph** (to model context) and incorporation of such context as features into SVM, Naive Bayes and Decision Tree **classifiers**.	**Semantic information can help to better understand** the **contextual variances** in which radicalisation terms are used when conveying 'radicalised meaning' vs. when not. Understanding such variances can help to create more accurate radicalisation detection methods.
(Fernandez et al., 2018)	Measure the **influence of online radicalisation** that a user is exposed to. Design a computational method based on the social science theory of roots of radicalisation (Schmid, 2013; Borum, 2016) (**user detection**)	17K tweets from pro-ISIS users and 122K tweets from 'general' **Twitter** users available via the Kaggle data science community (Kaggle, 2019)	Use **word vectors** to model the micro (individual), meso (social) and macro (global) radicalisation influence. **Cosine similarity** is used to compare such vectors against a Lexicon of radical terms	There is an important **need to leverage closer the knowledge of theoretical models** of radicalisation to design more effective technological solutions to track online radicalisation.
(Agarwal and Sureka 2015b)	Automatic identification of hate and extremism promoting tweets (**content detection**)	10,486 hate and terrorism-related **Twitter** posts (extracted based on hashtags) + 1M random tweets annotated by students for validation	They tested KNN and LibSVM **classifiers** based on religious, offensive, slang, negative emotions, punctuations and war related terms	Presence of **religious, war related terms, offensive words and negative emotions** are strong indicators of a tweet to be hate promoting
(Ashcroft et al., 2015)	Automatically detect messages released by jihadist groups on Twitter (**content detection**)	2,000 pro-ISIS **Twitter** posts (containing pro-ISIS terminology and extracted from the accounts 6,729 ISIS sympathisers), 2,000 anti-ISIS tweets(extracted from manually assessed anti-ISIS accounts), 2000 random tweets.[2]	Trained **classifiers** (SVM, Naive Bayes and Adaboost) based on stylometric (n-grams, hashtags, word frequency, etc.), time-based and sentiment features	Fridays are a key date to spread radical tweets. **Automatic detection is viable but can never replace human analysts.** It should be seen as a complementary way to detect radical content

Table 3: Approaches that focus on the **prediction** of online radicalisation



Work	**Goal**	**Data**	**AI Algorithm / Technique**	**Conclusions**
(Magdy et al., 2016)	Proposed an approach to **predict future support or opposition to ISIS**	57,000 **Twitter** users who authored or shared tweets mentioning ISIS. Categorised as pro or anti-ISIS based on the use of the full name of the group vs. an abbreviated form	SVM **classifier** based on bag-of-words features, including individual terms, hashtags, and user mentions	**Pro- and anti-ISIS users can be identified before they voice explicit support or opposition.**
(Ferrara et al., 2016)	Propose a computational framework for detection and **prediction of: adoption of radical content and interaction with pro-ISIS accounts**	Over 3M **Twitter** posts generated by over 25 thousand extremist accounts (manually identified, reported, and suspended by Twitter). 29M posts from the followers of these accounts	Random forest and logistic regression **classifiers** are used for classification and prediction based on user metadata and activity features, time features, and features based on network statistics	The ratio of retweets to tweets, the average number of hashtags adopted, the sheer number of tweets and the average number of retweets generated by each user, systematically rank very **high in terms of predictive power**

As the sections in this section addressing methods of evasion note, not all content is as explicit as this, and detecting extreme digital *speech* requires a different approach. In 2017, Google launched Perspective API, a machine-learning tool which uses natural language processing to score the perceived impact that a comment might have in a conversation. This technology is able to assign more disrespectful and aggressive comments a higher "toxicity" score, and thus aims to reduce the burden on human moderators by removing the need to manually remove every comment, instead focusing on high toxicity content. Platform regulation of comments poses a unique challenge, which at times may prove more difficult than regulating audio/visual content or posts. Programming and computer software scholars such as **Risch and Krestel (2020)** test and propose automated machine deep-learning, neural network, and sentiment analysis-driven solutions to strengthen platforms ability to combat toxic comments. **Maschechkin et al (2019)** propose original language-independent algorithms for pattern-based information retrieval, thematic modeling, prediction of message flow characteristics, and assessment and prediction of potential risk coming from members of online communities by using data on the structure of relations in these communities, allowing for the possibility of detecting dangerous users even without full access to the content they distribute (e.g. private channels and chatrooms). **Bérubé et al (2021)** recently argued that, while some moderation strategies targeted individuals most influential in the diffusion of unwanted material and others have focused on censorship of the content itself, few approaches consider both aspects simultaneously. Accordingly, they propose socio-semantic network analysis to identify individuals and subgroups strategically positioned in radical networks and whose comments encourage violence.[22]

In any case, it is apparent that the game of "cat and mouse" played between platforms and extremist organizations will endure for the foreseeable future, as new, primarily automated technologies strengthen platforms' ability to detect and remove malicious content and reduce the necessity for human moderation coincide with extremists' continual adoption of new methods to obfuscate those very technologies. The next section, "Contemporary Advancements in the Realm of Content Moderation," will address this in further detail.

[22] This discussion recounts the researchers' efforts in their own words, primarily to demonstrate that a substantial degree of work in computer science exists on the subject of anti-extremist content moderation, many of which propose novel solutions for platforms to employ.

Bibliography

Aliapoulios, Max, Bevensee, Emmi, Blackburn, Jeremy, Bradlyn, Barry, De Cristofaro, Emiliano, Stringhini, Gianluca, and Zannettou, Savvas. "An Early Look at the Parler Online Social Network." *Proceedings of the International AAAI Conference on Web and Social Media* 15, no. 1. 2021. https://arxiv.org/abs/2101.03820

Bérubé, Maxime, Beaulieu, Laurie-Anne, Mongeau, Pierre, and Saint-Charles, Johanne. "Identifying Key Players in Violent Extremist Networks: Using Socio-Semantic Network Analysis as Part of a Program on Content Moderation." *Studies in Conflict & Terrorism.* May 2021. https://www.tandfonline.com/doi/abs/10.1080/1057610X.2021.1927203

Bickert, Monika and Fishman, Brian. "Hard Questions: How Effective Is Technology in Keeping Terrorists off Facebook?" *Facebook*. April 23, 2018. https://about.fb.com/news/2018/04/keeping-terrorists-off-facebook/

Bindner, Laurence and Gluck, Raphael. "Jihadist Extremism" in *Extreme Digital Speech: Contexts, Responses and Solutions* ed. Bharath Ganesh and Jonathan Bright. *Vox-Pol*. 2020. https://www.voxpol.eu/download/vox-pol_publication/DCUJ770-VOX-Extreme-Digital-Speech.pdf

Bindner, Laurence and Gluck, Raphael. "Wilayat Internet: Islamic State's Resilience Across the Internet and Social Media." *Vox-Pol*. October 4, 2017. https://www.voxpol.eu/wilayat-internet-isis-resilience-across-internet-social-media/

Bodó, Loránd. "Decentralised Terrorism: The Next Big Step for the So-Called Islamic State (IS)? *Vox-Pol*. December 12, 2018. https://www.voxpol.eu/decentralised-terrorism-the-next-big-step-for-the-so-called-islamic-state-is/

Borelli, Marguerite. "Social Media Corporations as Actors of Counter-Terrorism." *New Media & Society*. August 2021. https://journals.sagepub.com/doi/abs/10.1177/14614448211035121

Chancellor, Stevie, Pater, Jessica, Clear, Trustin, Gilbert, Eric, and Choudhury, M.D. "#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities." *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 2016. https://www.semanticscholar.org/paper/%23thyghgapp%3A-Instagram-Content-Moderation-and-in-Chancellor-Pater/d90c289a539842a17c9a19e20142b3b50977273c

Chandrasekharan, Eshwar, Jhaver, Shagun, Bruckman, Amy, and Gilbert, Eric. "Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit." *Manuscript submitted to ACM*. September 2020. https://arxiv.org/abs/2009.11483

Clifford, Bennett and Powell, Helen. "Encrypted Extremism: Inside the English-Speaking Islamic State Ecosystem on Telegram." *Program on Extremism, The George Washington University*. 2019. https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EncryptedExtremism.pdf

Donovan, Joan, Lewis, Becca, and Friedberg, Brian. "Parallel Ports: Sociotechnical Change from the Alt-Right to Alt-Tech." *Post-Digital Cultures of the Far Right* ed. Maik Fielitz and Nick Thurston. 2019. https://www.degruyter.com/document/doi/10.14361/9783839446706-004/html

Fernandez, Miriam and Alani, Harith. "Artificial Intelligence and Online Extremism: Challenges and Opportunities." in *Predictive Policing and Artificial Intelligence* ed. John McDaniel and Ken Pease. 2021. http://oro.open.ac.uk/69799/

Ferreira, Ricardo Ribeiro. "Liquid Disinformation Tactics: Overcoming Social Media Countermeasures through Misleading Content." *Journalism Practice*. April 2021. https://doi.org/10.1080/17512786.2021.1914707

Gallacher, John. "Automated Detection of Terrorist and Extremist Content" in *Extreme Digital Speech: Contexts, Responses and Solutions* ed. Bharath Ganesh and Jonathan Bright. *Vox-Pol*. 2020. https://www.voxpol.eu/download/vox-pol_publication/DCUJ770-VOX-Extreme-Digital-Speech.pdf

Gibson, Anna. "Free Space and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces." *Social Media & Society.* March 2019. https://journals.sagepub.com/doi/full/10.1177/2056305119832588

Gerrard, Ysabel. "Beyond the Hashtag: Circumventing Content Moderation on Social Media." *New Media & Society*. May 2018. https://journals.sagepub.com/doi/full/10.1177/1461444818776611?casa_token=57nBGNDDb2MAAAAA%3ARjBUu5iWpc8JUAuDg4d6TTmLLF4HTTJfsI8uq0eBZxT89PIBOlqw49ZVpvZ5Oa-BmQJ46DtR1sI

Glaser, April. "White Supremacists Still Have a Safe Space Online." *Slate*. October 9, 2018. https://slate.com/technology/2018/10/discord-safe-space-white-supremacists.html

Gröndahl, Tommi, Pajola, Luca, Juuti, Mika, Conti, Mauro, and Asokan, N. "All You Need is 'Love': Evading Hate Speech Detection." *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. October 2018. https://dl.acm.org/doi/abs/10.1145/3270101.3270103?casa_token=C02P838RnugAAAAA:tSEsEqaQ6mayMsDq8ZEVhyJSV8rZcrhn9XRTRjlk6dVVFGW__hLD3XizDHT_ulJMq-wmy6GLp-Q

Gray, Joanne E., and Suzor, Nicolas P. "Playing with Machines: Using Machine Learning to Understand Automated Copyright Enforcement at Scale." *Big Data & Society*. April 2020. https://journals.sagepub.com/doi/full/10.1177/2053951720919963

Heller, Brittan. "Combating Terrorist-Related Content Through AI and Information Sharing." *Transatlantic Working Group*. 2019. https://cdn.annenbergpublicpolicycenter.org

Jardine, Eric. "Online Content Moderation and the Dark Web: Policy Responses to Radicalizing Hate Speech and Malicious Content on the Darknet." *First Monday* 24, no. 12. December 2019. https://firstmonday.org/ojs/index.php/fm/article/view/10266

Jasser, Greta, McSwiney, Jordan, Pertwee, Ed, and Zannettou, Savvas. "'Welcome to #GabFam': Far-Right Virtual Community on Gab." *New Media & Society*. June 2021. https://journals.sagepub.com/doi/abs/10.1177/14614448211024546

Jhaver, Shagun. "Identifying Opportunities to Improve Content Moderation." *Dissertation, Georgia Institute of Technology.* 2020. https://smartech.gatech.edu/handle/1853/62779

Jhaver, Shagun, Appling, Darren Scott, Gilbert, Eric, and Bruckman, Amy. "'Did you Suspect the Post Would Be Removed?': Understanding User Reactions to Content Removals on Reddit." *Proceedings of the ACM on Human-Computer Interaction.* November 2019. https://dl.acm.org/doi/abs/10.1145/3359294

Jhaver, Shagun, Boylston, Christian, Yang, Diyi, and Bruckman, Amy. "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter." *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, no. 381. October 2021. https://shagunjhaver.com/files/research/jhaver-2021-deplatforming.pdf

Jhaver, Shagun, Birman, Iris, Gilbert, Eric, and Bruckman, Amy. "Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator." *ACM Transactions on Computer-Human Interaction* 26, no. 5, article 31. July 2019. https://dl.acm.org/doi/10.1145/3338243

Jhaver, Shagun, Ghoshal, Sucheta, Bruckman, Amy, and Gilbert, Eric. "Online Harassment and Content Moderation: The Case of Blocklists." *ACM Transactions on Computer-Human Interaction* 25, no. 2. April 2018. https://dl.acm.org/doi/abs/10.1145/3185593

Jiang, Jialun Aaron, Kiene, Charles, Middler, Skyler, Brubaker, Jed R., and Fiesler, Casey. "Moderation Challenges in Voice-Based Online Communities on Discord." *Proceedings of the ACM on Human-Computer Interaction*. November 2019. https://arxiv.org/abs/2101.05258

Kiesler, Sara, Kraut, Robert E., Resnick, Paul, and Kittur, Aniket. "Regulating Behavior in Online Communities" in *Building Successful Online Communities: Evidence-Based Social Design* ed. Robert E. Kraut and Paul Resnick. *MIT Press*. 2012. https://books.google.com/books?hl=en&lr=&id=lIvBMYVxWJYC&oi=fnd&pg=PA125&dq=kiesler+2012&ots=z_B1drm7HE&sig=xZY-N4Hv8XgwgEJimX0pad6JN_w#v=onepage&q=kiesler%202012&f=false

Lima, Lucas, Reis, Julio C.S., Melo, Philipe, Murai, Fabricio, Araujo, Leandro, Vikatos, Pantelis, and Benevenuto, Fabricio. "Inside the Right-Leaning Echo Chambers: Characterizing Gab, and Unmoderated Social System." *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2018. https://ieeexplore.ieee.org/abstract/document/8508809

Llansó, Emma, van Hoboken, Joris, Leerssen, Paddy, and Harambam, Jaron. "Artificial Intelligence, Content Moderation, and Freedom of Expression." *Transatlantic Working Group*. February 26, 2020. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

Lyons, Henrietta, Velloso, Eduardo, and Miller, Tim. "Conceptualizing Contestability: Perspectives on Contesting Algorithmic Decisions." *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, no. 106. 2021. https://dl.acm.org/doi/abs/10.1145/3449180

Lyons, Henrietta, Velloso, Eduardo, and Miller, Tim. "Designing for Contestation: Insights from Administrative Law." *Proceedings of the 2019 CSCW Workshop on Contestability in Algorithmic Systems*. 2021. https://arxiv.org/abs/2102.04559

Lyons, Henrietta, Velloso, Eduardo, and Miller, Tim. "Fair and Responsible AI: A Focus on the Ability to Contest." *Fair and Responsible AI Workshop*. April 2020. https://arxiv.org/abs/2102.10787

Mashechkin, I.V., Petrovskiy, M.I., Tsarev, D.V., and Chikunov, M.N. "Machine Learning Methods for Detecting and Monitoring Extremist Information on the Internet." *Programming and Computer Software* 45. 2019. https://link.springer.com/article/10.1134/S0361768819030058

Matias, J. Nathan. "Preventing Harassment and Increasing Group Participation Through Social Norms in 2,190 Online Science Discussions." *Psychological and Cognitive Sciences*. May 2019. https://www.pnas.org/content/116/20/9785

Risch, Julian and Krestel, Ralf. "Toxic Comment Detection in Online Discussions." *Deep Learning-Based Approached for Sentiment Analysis*. January 2020. https://link.springer.com/chapter/10.1007/978-981-15-1216-2_4

Singh, Spandana. "Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content." *New America: Open Technology Institute*. July 22, 2019. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

Singh, Shubham, Kaushal, Rishabh, Buduru, Arun Balaji, and Kumaraguru, Ponnurangam. "KidsGUARD: Fine Grained Approach for Child Unsafe Video Representation and Detection." *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. April 2019. https://dl.acm.org/doi/abs/10.1145/3297280.3297487

Srinivasan, Kumar Bhargav, Danescu-Niculescu-Mizil, Cristian, Lee, Lillian, and Tan, Chenhao. "Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community." *Proceedings of the ACM on Human-Computer Interaction*. November 2019. https://dl.acm.org/doi/abs/10.1145/3359265

Suzor, Nicolas. "Understanding Content Moderation Systems: New Methods to Understand Internet Governance at Scale, Over Time, and Across Platforms." *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* ed. Ryan Whalen. 2020. https://www.elgaronline.com/view/edcoll/9781788977449/9781788977449.00013.xml

Suzor, Nicolas P., West, Sarah Myers, Quodling, Andrew, and York, Jillian. "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation." *International Journal of Communication* 13. 2019. https://ijoc.org/index.php/ijoc/article/view/9736

Vaccaro, Kristen. Sandvig, Christian, and Karahalios, Karrie. "'At the End of the Day Facebook

Does What it Wants': How Users Experience Contesting Algorithmic Content Moderation." *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2, no. 167. 2020. https://dl.acm.org/doi/abs/10.1145/3415238

van der Vegt, Isabelle, Gill, Paul, Macdonald, Stuart, and Kleinberg, Bennett. "Shedding Light on Terrorist and Extremist Content Removal." *Global Research Network on Terrorism and Technology: Paper No. 3.* 2019. https://rusi.org/explore-our-research/publications/special-resources/shedding-light-on-terrorist-and-extremist-content-removal

West, Sarah Myers. "Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms. " *New Media & Society*. 2018. https://journals.sagepub.com/doi/abs/10.1177/1461444818773059

West, Sarah Myers. "Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms." *Media and Communication* 5, no. 3. 2017. https://www.cogitatiopress.com/mediaandcommunication/article/view/989

Zannettou, Savvas, Bradlyn, Barry, De Cristofaro, Emiliano, Kwak, Haewoon, Sirivianos, Michael, Stringini, Gianluca, and Blackburn, Jeremy. "What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber." *Proceedings of the The Web Conference 2018*. 2018. https://dl.acm.org/doi/abs/10.1145/3184558.3191531

Zhu, Wanzheng, Gong, Hongyu, Bansal, Rohan, Weinberg, Zachary, Christin, Nicolas, Fanti, Giulia, and Bhat, Suma. "Self-Supervised Euphemism Detection and Identification for Content Moderation." *IEEE Symposium on Security & Privacy*. 2021. https://arxiv.org/abs/2103.16808

8. Contemporary Advancements in Online Content Moderation

The following section provides a more opinionated overview of the literature.
Prepared primarily by LB

Background

Over the last half decade—and especially over the course of the last two years—major social media and online platforms have integrated automated technologies (AI) into the content moderation process. The pace of this transition was exacerbated by the COVID-19 Pandemic, which forced many major social media platforms to send their human moderators home. Apart from the COVID-19 Pandemic, several developments (e.g. the sheer volume of content published on platforms today, an uptick in misinformation campaigns, ever-changing obfuscation strategies, the need to outsource commercial content moderation (CCM) work, and the well-documented effects that CCM labor has on workers) have necessitated this AI integration. Building off the previous section's examination of user reactions to moderation, common methods of evasion, and the countermeasures platforms take to combat such evasiveness, the present section provides a background and literature review of contemporary technological advancements in the realm of online content moderation. Some articles outline methods platforms have already developed and employed, or provide relevant background information; the bulk, however, propose and test various novel methods (e.g. **machine learning (ML)**, **pattern matching**, **natural language processing (NLP)**, **deep learning**, **sentiment analysis**, **service-based analysis**, **neural networks**). Accordingly, a much greater proportion of the articles cited in the present section are technical papers from computer scientists (often presented at/published by the Association for Computing Machinery, ACM). Thus, this section may be weaker than others with respect to its critical concept analysis, given my insufficient conceptual understanding of computer programming. Per usual, the order in which information appears does not signify pertinence, and I have included a bibliography with links to access the articles referenced at the conclusion. The early pages of the present section primarily provide relevant contextual and background information—often pulling from previous sections—whereas the newest articles (and those most pertinent to the topic at hand) can be found in the *Potential New Tools* and *Automated Moderation in Practice Today* sections in the latter half.

At the outset, it is worth highlighting the great obstacle obstructing scholars seeking to understand how platforms presently do moderate and how effective those strategies are—platforms themselves. As eminent content moderation scholar Sarah T. Roberts noted, "platforms operate their moderation practices under a complex web of nebulous rules and procedural opacity" and "are extremely reluctant to acknowledge, much less describe in great detail the internal policies and practices that govern what goes into the content ultimately available on the mainstream site. **(Roberts 2018)**. Many platforms do publish periodic transparency reports—which typically disclose aggregate, quantitative data overviews about requests for content removal, rates of compliance on the platform, and sometimes the frequency of user appeal—but these have major limitations, even for more thorough reports such as the Community Guidelines enforcement reports published by Facebook, Twitter, and YouTube for the first time in 2018. For one, the aggregated data only shows the platforms' assessments, not the merits, inhibiting researchers from evaluating the accuracy of decisions. Further, transparency reports vary widely from company to company with respect to how they classify data and how much detail they provide.[1] Accordingly, researchers needing to know what content was actually removed often turn to the Lumen database, which archives legal takedown notices from any platform or sender that chooses to share them; to date, it is the best source for such information, holding some 9.3M notices targeting approximately 3.35B URLs, mostly from Google. Scholars have also turned to issue-specific platform disclosures, such as Google's 2018 report *Three Years of the Right to Be Forgotten*, as well as public filings, other government disclosures, and published reports from independent auditors such as The Global Network Initiative **(Keller and Leerssen 2020)**. **Anderson, Stender, Myers West, and York (2016)** turned to crowdsourced/solicited data gathered over the course of four months from a questionnaire on the onlinecensorship.org website, aggregating and analyzing the data across geography, platform, content type, and issue areas to highlight trends in social media censorship and variation between platforms. In other words, aggregate, comparable data on platform moderation decisions can be so difficult to come by that even highly respected scholars have thrown major selection bias concerns to the wind just to have *something* to work with.

[1] For a comprehensive overview of transparency reporting, see **Singh and Bankston (2018),** who issued a report on the practice on behalf of New America's Open Technology Institute.

The Inadequacy of Human Moderation

Before AI quickly became the central tenet of the broader online content moderation discussion, scholars often focused on “commercial content moderation (CCM)” workers and firms, whose services platforms use to screen content. In several articles **(Roberts 2014; 2016a; 2016b; 2017)** and a subsequent book **(Roberts 2019)**, Sarah T. Roberts studied the politics and policies of platforms by exploring how the unseen efforts of CCM workers affect how users experience social media and user-generated content and what effect that work has on those workers themselves. In doing so, she builds off of prior work from content moderation scholar Tarleton Gillespie, who had studied CCM workers but generally focused on platforms themselves **(Gillespie 2010; 2015; 2018)**. Roberts employs a qualitative, ethnographic method centered around interviews with moderators, foreign and domestic, which illuminated a handful of findings. For one, human-led moderation efforts undertaken by large platforms were generally outsourced—in a manner remarkably similar to call centers—predominately to non-native English speakers in India. More poignantly, she found that tasks associated with such work range from mind-numbingly repetitive at best (workers, on average, spent < 2s deciding whether an image violated community guidelines) to psychologically damaging at works, and underpaid in any case.

Apart from CCM work, there does exist another human side to moderation—volunteer moderation. Particularly in the early years of social media, not all platforms employed human content moderators, instead relying on volunteer moderators from the site. This more primitive form of moderation—embodied by Reddit, Wikipedia, and AOL at the time, and by Gab and Discord today—relies on the custodial content work of unpaid moderators to create, support, and control public discourse for millions. **Matias (2019)** asked what this work meant, and who it was for, coupling a data collection from over 52,000 subreddits and interviews with volunteer moderators, determining that the behavior operates often as a form of “civic labor” akin to journalists who fact-check news on Facebook. A decade earlier, **Postigo (2009)** asked a similar question with respect to AOL’s extensive use of volunteers and the Dept. of Labor investigation which followed, arguing based on interviews and DOL documents (qualitatively) that the success of a co-productive relationship between platform and volunteer exists as a function of a balance between 1) the perceived reasonable compensation of volunteers, 2) social factors such as sense of community, and 3) a sense of accomplishment.

The Movement Toward Automated Moderation

Whereas bulletin boards and forums were, in the early internet, meticulously managed by these dedicated administrators within a community, platforms over the last decade began operating at a scale which has led them away from such practices. Soon thereafter, even the use of full-time, employed human moderators could not adequately monitor the volume of content produced. As **Gillespie (2020)** stated, "the quantity, velocity, and variety of content is stratospheric... the consequences of online harms now extend beyond the platform on which they occur; and criticism of the platforms and their failures to moderate has exploded, catalyzed by Gamergate, Myanmar, revenge porn, the 2016 U.S. Presidential Election, Alex Jones, and Christchurch." This emphasizes a key point; the push towards AI in online content moderation and the adoption of automated tools by major platforms predates the Pandemic and even perhaps the scale/size issues which necessitated it.

YouTube began adopting increasingly advanced and ambitious moderation technologies well before the 2016 U.S. Election controversies, particularly in response to the surge in extremist content which paralleled the rise of the Islamic State (IS).   For example, ISIS fighters created the "Syrian Archive," a software that would download new material from Syrian YouTube channels to its servers before they were purged from the site, forcing the platform to ramp up moderation to an extent at which it could delete videos at a speed faster than the Archive could download them. A challenge of this scale and urgency necessitated the use of AI, which brings about its own problems, such as mistaken deletions or blockings, or ones that activists or governments disagree with.[2] YouTube reported that their machine learning algorithms helped remove more than 150K videos depicting violent extremism form June to December 2017 alone, equivalent to the work of ~180K full time moderators **(Chen 2021).**

Since, several major platforms have adopted some form of scraper—operating in conjunction with a myriad of other strategies—with varying degrees of transparency regarding

[2] For more on YouTube's efforts to combat extremist content and statistical analyses of the Scraper's effectiveness, consult the prior two sections and/or Suzor (2020) and Gray and Suzor (2020).

its functionality. Extremism—particularly jihadism—has diminished relative to other content moderation concerns in the last half decade, due in large part to platforms' newly aggressive countermeasures and the subsequent migration of extremist users and groups to unmoderated and/or alternative communications platforms (e.g. Gab, Telegram, Parler, Discord, the DarkWeb). Nonetheless, some jihadists continue to operate to some degree on the surface web and even popular platforms—distorting images, cropping, hijacking hashtags, hiding media in multiple uploads in a single post, and using disappearing stories—in a game of cat-and-mouse which perpetually incentivizes novel technological advancements in the realm of moderation.[3]

Facebook, now boasting 2B users—from whom 4K are photos are uploaded every second, and more than 4B video views take place every day. It is no surprise that in May 2017, Mark Zuckerberg said that Facebook receives millions of complaints per week. YouTube reported that 500 hours of videos are uploaded every minute as of May 2019.

In a recent policy brief for the AI4Dignity project, **Udupa et al (2021)** overviewed the challenges (more sociological than technical) facing AI in content moderation as follows:

i. **Inability to fully account for evolving context and practice**—AI lacks the ability to understand the context, intent, linguistic nuances, cultural variation, and the changes that occur around some of these required to evaluate content. It can also be difficult for AI to account for misspelled words, changing syntax, and the use of images, GIFs, and memes to convey offense.

ii. **Biased/inaccurate/limited training data**—AI can be trained on limited, skewed, or biased data sets, resulting in decisions that are inaccurate and/or reproduce these limitations. It can also result in decisions that are inconsistent across contexts, raising concerns of potentially discriminatory content moderation practices and the removal of legitimate content, particularly with respect to speech from minority and historically disadvantaged groups.

iii. **Linguistic Diversity**—Though companies are continuing to invest in natural language processing (NLP) models that cover several major global languages (e.g. English, Spanish, and Mandarin), smaller languages and those spoken in poorer countries have been left behind except in places where international outcry has increased pressure (e.g. Myanmar). This lack of diversity not only disadvantages such people, but also can result in extreme speech being unidentified or misidentified.

[3] For a thorough look at methods of evading content moderation and the countermeasures platforms adopt, consult the previous section.

iv. **Function creep**—Relying predominately on AI to monitor and moderate content online can result in "function creep," "the spilling over of technologies devised for certain purposes into other areas, uses, purposes, with impacts on safety, privacy and bias." In other words, user privacy concerns.

v. **Opaque decisions and lack of transparency**—A lack of transparency around how AI is developed and trained for content moderation purposes can result in opaque decisions escaping public scrutiny.

vi. **Lack of notice**—A lack of notice on when and how AI is being used in decisions pertaining to content limits the ability for users to appeal the decision.

vii. **Proactive moderation**—The use of AI to moderate content prior to publication can result in censorship and a alack of due process for the users as they may not be aware that their content was filtered/restricted or have the ability to appeal the decision.

viii. **Shifting the burden of determining unlawful content**—Legal requirements for companies to use AI in moderating content shifts the burden of determining legality to companies and removes important safeguards such as judicial review.

ix. **Authoritarian use**—Governments can use AI to facilitate or mandate authoritarian or unlawful censorship practices.

x. **Amplification of harmful content**—When AI is used in systems that recommend and prioritize content, this can amplify problematic content depending on the characteristics that the AI has been trained to prioritize.

The authors also outline areas for further reflection and development with respect to their "people-centric" proposed approaches to AI-assisted content moderation, and highlight the limitations, challenges, and recommendations for future development (particularly from the academic side of things), some of which I've included below:

i. **Sparsity of data and interpretative labor**: Involving communities for context-specific annotations is resource intensive; whereas NLP researchers have often used automated collection methods from easy access sources (e.g. MTurk) to gather large volumes of data. Thus, datasets which are more representative of actual vitriolic comments directed at different target groups are smaller and harder to gather.

ii. **Computational costs and internet access**: In involving diverse communities from regions with vastly different levels of internet access and technological resources, it is important to not only take into account the performance of ML models but also the size of the model and prediction time; in other words, the tool should be able to work on low-end devices, and safeguards should be put into place to ensure data is not lost in cases of internet disconnection.

iii. **Open models for data annotation**: Another broad challenge is to build on existing efforts to develop open models for annotation that can incorporate human input at scale. Since training models in NLP are often beyond what academic projects can afford, the challenge is to make use of publicly available, open-source models as a starting point for further training.

iv. **Collaborative graded approach**: Further emphasis should be placed on developing principles that could guide a graded approach to human-AI collaboration, such as type of content, nature of content, and potential of harm, to determine the extent to which the AI augments a decision vs. takes a decision.

v. **Linguistic and cultural diversity:** Future development of collaborative models will depend on involving communities (e.g. fact checkers) from diverse linguistic and cultural background, since current NLP models are heavily tilted towards large, resource rich languages and linguistic communities.

vi. **Multimodal content**: Online extreme speech that combines moving images, text and audio—exemplified by internet memes and GIFs—poses specific challenges to ML models.

Automated Moderation in Practice Today

Just as extremist groups purveying malicious content have continually evolved in their capacity to obfuscate moderation, platforms have continually evolved in their capacity to detect and remove unwanted material. As with any subtopic under the umbrella of content moderation, advancements in AI and automated detection technology are pertinent to counterterrorism on the internet. The primary automated tools used to curate, organize, filter, and classify information in the online content moderation process are as follows **(Singh 2019)**:

i. **Digital Hash Technology**—which works by converting images and videos from an existing database into grayscale format, overlaying them onto a grid, and assigning each square a numerical value. The designation of a numerical value converts the square into a hash, or digital signature, which remains tied to the image or video and can be used to identify other iterations of the content either during ex-ante moderation or ex-post proactive moderation. Digital hash technology has been widely adopted by internet platforms. Child Sexual Abuse Material (CSAM) detection technology, known as PhotoDNA, a hash technology, has expanded to become a powerful tool, even creating newer technologies like ContentID, a

YouTube technology which allows users to create digital hashes for the content to protect against copyright violations. Despite its power and apparent effectiveness, several concerns have arisen (elaborated upon further in this section), such as the definition of extremist content, bias in algorithmic training (e.g. reduced reliability of automated tools trained to focus on IS or al-Qaeda in addressing the larger corpus of extremist content), and the lack of transparency and accountability around how digital hash technology is deployed.

ii. **Image Recognition**—which is employed by digital hash technologies, but can also be used more broadly in the moderation process. During ex-post proactive moderation, for instance, image recognition tools can identify specific objects within an image (e.g. weapons) and decide bsed on factors including user experience and risk whether the image should be flagged to a human for review. Such automated recognition tools are currently employed by platforms to filter through and prioritize cases for CCM workers. Concerns exist surrounding such technologies' frequent inability to incorporate more nuanced and contextual insights into their detection procedures, the quality of datasets models are trained on, and the lack of transparency around how databases are compiled, what types of content they focus on, how accurate they are across different categories of content, and how much user expression has been accurately and erroneously removed as a result.

iii. **Metadata Filtering**—which utilizes files' metadata, the descriptive characteristics about content. Such filtering tools can be used during ex-ante and ex-post proactive moderation to search a series of filed to identify content that fits a particular set of metadata parameters. This tool is used particularly to identify copyright-infringing materials. However, because metadata manipulation and mislabeling can be easy, the effectiveness and accuracy of such tools are limited, particularly compared to the others.

iv. **Natural Language Processing (NLP)**—a set of techniques that uses computers to parse text; in the context of content moderation, text is typically parsed in order to make predictions about the meaning of the text, such as what sentiments it indicates. NLP classifiers are particularly used to detect hate speech and extremist content, and to perform sentiment analysis. NLP classifiers are generally trained on text examples that have been annotated by humans in order to indicate whether they belong to a particular category (e.g. extremist v. non-extremist). Though platforms have been increasingly exploring and adopting NLP classifiers, the technology is still limited for several reasons: a) NLP technologies are domain-specific, which means they can only focus on one particular type of objectionable content; b) because there is significant variation in how speech is expressed, these categories are narrow; c) finding and compiling comprehensive enough datasets to train NLP classifiers is challenging, expensive, and tedious; d) in order for NLP classifiers to operate

> accurately, they need to be provided with clear and consistent parameters and definitions of speech.[4]

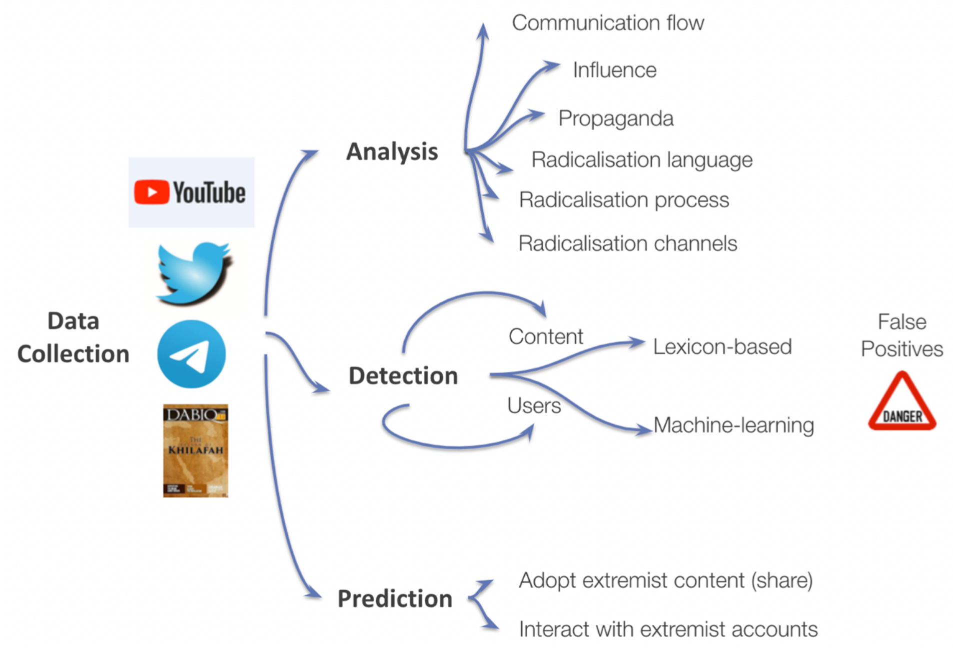
In order to combat migration of extremists from one platform to another, technology firms (initially Google, Microsoft, Facebook, and Twitter) set up the Global Internet Forum to Counter Terrorism (GIFCT) to create a shared database of extremist content to block repeated uploads of the same material. They do so by creating a cryptographic hash (i.e. a digital fingerprint) of each identified image or video clip, automatically blocking database matches that are found when content is uploaded to any of those platforms. As of 2019, the database contained 88K digital fingerprints, and smaller platforms like Instagram, LinkedIn, and Snapchat have been added. In a report for the Transatlantic Working Group, **Heller (2019)** concedes that early data indicate that the GIFCT's hash-sharing efforts are working, at least based on what information we do have, and if content removal is the metric of success. Statistics from the GIFCT seem to reinforce its claim that it is effective at removing a small but significant slice of content which accounts for the worst types of terror-related content. Following a series of interviews with GIFCT partner companies and law enforcement, **van der Vegt et al (2019)** similarly argue that this database of hashes—focusing on the most extreme and egregious terrorist images and videos—should be expanded to include more borderline content.

The UK government, in partnership with the AI firm ASI Data Science, announced a similar approach in 2018, employing machine learning to analyze a range of subtle patterns within videos and determine whether they match patterns found in IS propaganda. Able to distinguish original content from videos that discuss the same imagery in different contexts such as news reporting, this tool is reportedly 94% accurate, with a false positive rate of 0.005%. Such a tool attests to the power AI offers in the realm of online counterterrorism, as such a tool would take approximately 20,000 people to perform the same tasks; nonetheless, the program does not remove content on its own, but does flag videos for human consideration **(Gallacher 2019).** Platforms themselves have found automated detection to be immensely valuable as a tool; speaking on behalf of Facebook, **Bickert and Fishman (2018)** note that 99% of IS content was removed before reported to the platform by a user. Similarly, YouTube reported on their official

[4] I highly recommend exploring this thorough report from **Singh (2019)**, "Everything in Moderation: An Analysis of How Internet Platforms are Using Artificial Intelligence to Moderate User-Generated Content."

blog that 98% of the videos removed for violating violent extremism rules are detecting by an automated system. **Gillespie (2020)** argues that such statistics touted by platforms are misleading, however, as they imply that machine learning (ML) techniques are accurately spotting new instances of abhorrent content, when at least for now, the overwhelming majority of what is being automatically identified are copies of content that have already been reviewed by a human moderator.

As the sections in this section addressing methods of evasion note, not all content is as explicit as this, and detecting extreme digital *speech* requires a different approach. In 2017, Google launched Perspective API, a machine-learning tool which uses natural language processing to score the perceived impact that a comment might have in a conversation. This technology is able to assign more disrespectful and aggressive comments a higher "toxicity" score, and thus aims to reduce the burden on human moderators by removing the need to manually remove every comment, instead focusing on high toxicity content. Platform regulation of comments poses a unique challenge, which at times may prove more difficult than regulating audio/visual content or posts. **Fernandez and Alani (2021)** produce the following flow chart overviewing AI approaches to counter online radicalization:



Meanwhile, **Gorwa, Binns, and Katzenbach (2020)** break down the typological and function of various algorithmic moderation systems, which platforms employ them and to what end, and the human role within these systems, in the tables pasted below:

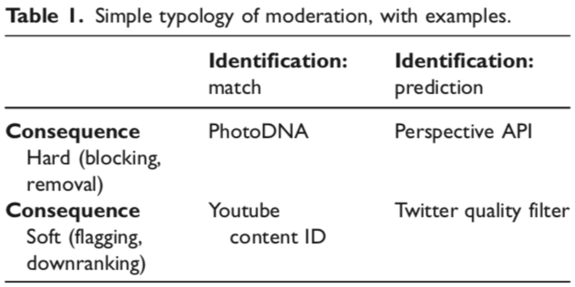


Table 1. Simple typology of moderation, with examples.

	Identification: match	**Identification:** prediction
Consequence Hard (blocking, removal)	PhotoDNA	Perspective API
Consequence Soft (flagging, downranking)	Youtube content ID	Twitter quality filter

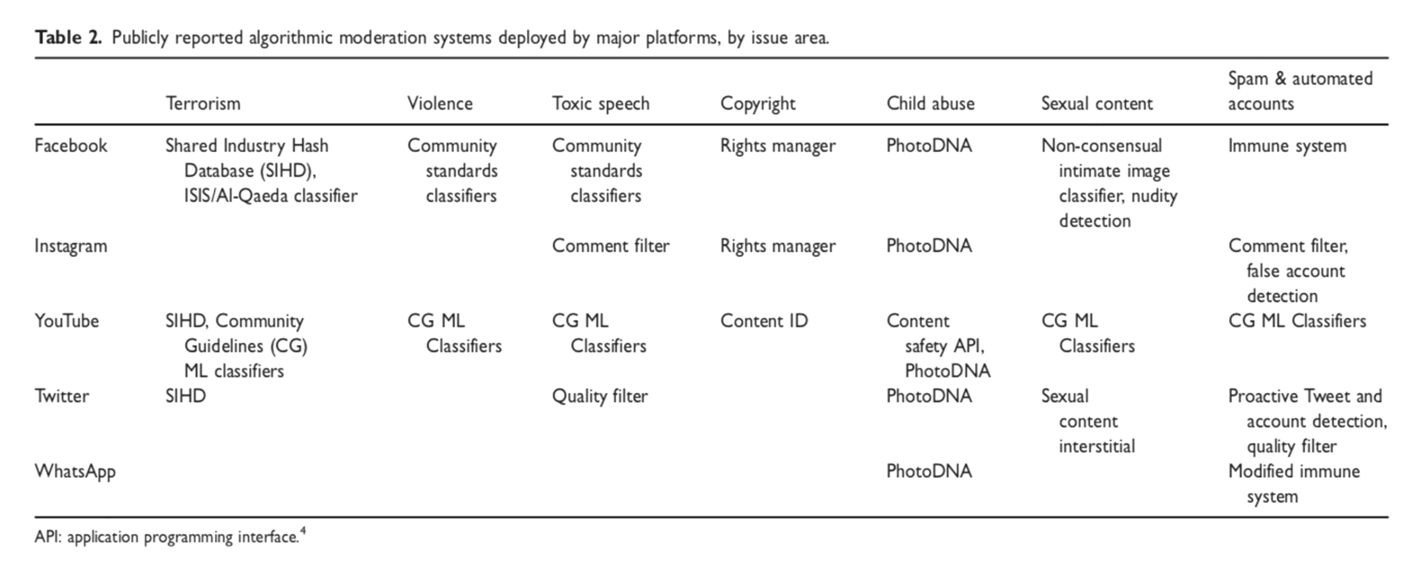


Table 2. Publicly reported algorithmic moderation systems deployed by major platforms, by issue area.

	Terrorism	Violence	Toxic speech	Copyright	Child abuse	Sexual content	Spam & automated accounts
Facebook	Shared Industry Hash Database (SIHD), ISIS/Al-Qaeda classifier	Community standards classifiers	Community standards classifiers	Rights manager	PhotoDNA	Non-consensual intimate image classifier, nudity detection	Immune system
Instagram			Comment filter	Rights manager	PhotoDNA		Comment filter, false account detection
YouTube	SIHD, Community Guidelines (CG) ML classifiers	CG ML Classifiers	CG ML Classifiers	Content ID	Content safety API, PhotoDNA	CG ML Classifiers	CG ML Classifiers
Twitter	SIHD		Quality filter		PhotoDNA	Sexual content interstitial	Proactive Tweet and account detection, quality filter
WhatsApp					PhotoDNA		Modified immune system

API: application programming interface.[4]

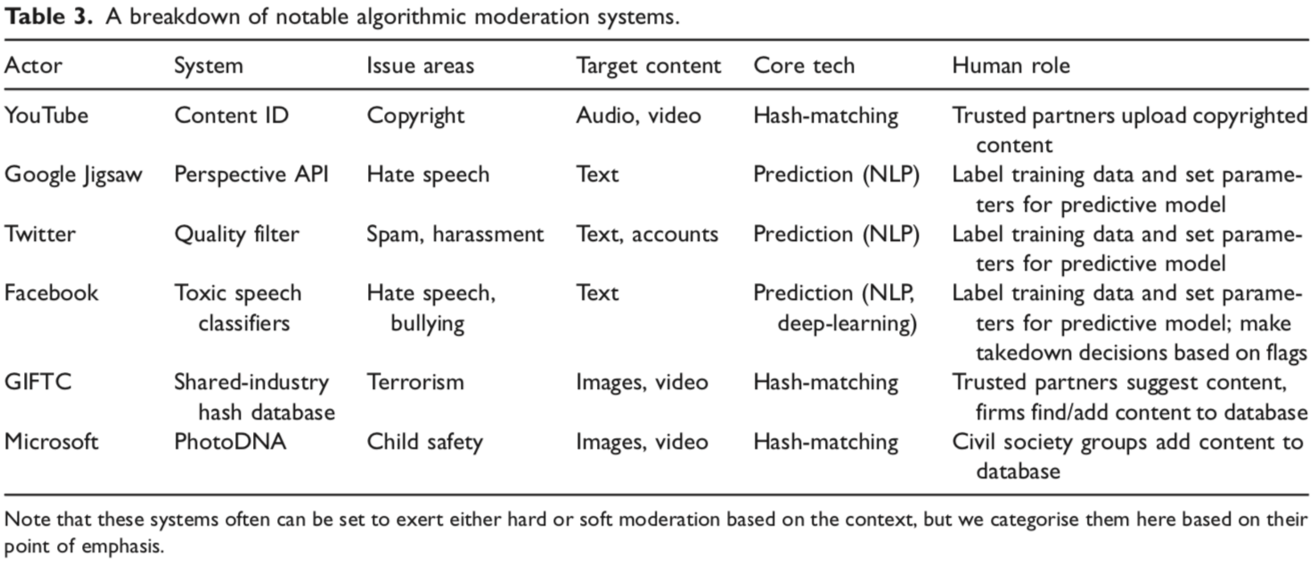


Table 3. A breakdown of notable algorithmic moderation systems.

Actor	System	Issue areas	Target content	Core tech	Human role
YouTube	Content ID	Copyright	Audio, video	Hash-matching	Trusted partners upload copyrighted content
Google Jigsaw	Perspective API	Hate speech	Text	Prediction (NLP)	Label training data and set parameters for predictive model
Twitter	Quality filter	Spam, harassment	Text, accounts	Prediction (NLP)	Label training data and set parameters for predictive model
Facebook	Toxic speech classifiers	Hate speech, bullying	Text	Prediction (NLP, deep-learning)	Label training data and set parameters for predictive model; make takedown decisions based on flags
GIFTC	Shared-industry hash database	Terrorism	Images, video	Hash-matching	Trusted partners suggest content, firms find/add content to database
Microsoft	PhotoDNA	Child safety	Images, video	Hash-matching	Civil society groups add content to database

Note that these systems often can be set to exert either hard or soft moderation based on the context, but we categorise them here based on their point of emphasis.

As Gallacher notes**,** challenges for platforms still exist: a) how to deal with false positives; b) how to prevent malicious attacks on the tools themselves; c) how to apply these tools when groups move to more encrypted spaces and smaller platforms; and d) how to appropriately apply tools built with a focus on one domain of extremism to other forms. **Fernandez and Alani (2021)** frame the challenges facing AI in content moderation a bit differently, as follows: a) the lack of a common definition of prohibited and extremist internet activity; b) the lack of solid verification of the datasets collected to develop detection and prediction models; c) the lack of cooperation across research fields, since most of the developed technological solutions are neither based on, nor do they take advantage of existing social theories and studies of radicalization; d) the constant evolution of behaviors associated with online extremism in order to avoid being detected by the developed algorithms; and e) the development of ethical guidelines and legislation to regulate the design and development of AI technology to counter radicalization.

Common challenges for toxic comment classification among different datasets arise from a variety of sources, such as: misclassification of comments, toxicity without swear words, non-toxicity with swear words; quotations, references, metaphors, and comparisons; sarcasm, irony, and rhetorical questions; mislabeled comments; and idiosyncratic and rare words (e.g. typos, slang, abbreviations, and neologisms).

In their comprehensive analysis of toxic comment detection in online discussions, **Risch and Krestel (2020)** codify various subtypes of toxicity, present various data sources and deep learning approaches tailored to sentiment analysis, and propose new augmentations for moderation systems. They categorize toxicity into the following subtypes, which are not mutually exclusive: obscene language/profanity; insults; threats; hate speech/identity hate; and otherwise toxic. Deep learning for sentiment analysis and, in particular, toxic comment classification, is mainly based on two pillars: large datasets and complex neural networks. Risch and Krestel argue that datasets often generate an inherent bias reflected in the programs (e.g. scrapers) they train, as data labeling is required for semi-supervised and super-vised learning approaches, and those annotations are determined by pre-defined lists of abusive language. Ultimately, since these data sets are collected with a focus on toxic comments, they overrepresent toxic content. Large datasets of toxic comments also allow for the training of complex neural networks with millions of parameters. Word embeddings are the basis of neural networks when working with text data in general and specifically in the case of toxic comment classification. They translate each word to a vector of 50-300 floating-point numbers, create dense vectors that can capture and represent word similarity. Neural networks for toxic comment classification either use recurrent neural network (RNN) layers, such as long short-term sectionry (LTSM) or gated recurrent unit (GRU) layers. In any case, today's industrial applications so far refrain from using deep learning models for content moderation due to the lack of explainability.

In many real-world applications, toxic comment classification is used to support a binary decision-making process: does a particular comment need moderation or can it be published right away? By offering a classification of toxic comment subtypes, the authors strive to distinguish between merely bad comments and criminal offenses, ultimately helping moderators make a final decision. Binary classification systems, they argue, oversimplify the different nuances in language, whereas a more fine-grained classification would give insights into why a comment is not suitable for publication. The application of such an approach would be difficult (as binary classification itself already is); until such a time when improvements can be made in explaining the decisions of deep neural networks, the authors argue, the industry will fall back to less complex models (e.g. logistic regression models which can explain which features make a comment inappropriate in a specific context).

Potential Tools

As early as 2010 (and almost certainly before) scholars have proposed and tested novel methods for moderating online content. As mentioned earlier, the primary concern underlying online content when much of the early work on algorithmic moderation was published was pro-IS content, and extremism more broadly; this is reflected in the literature of those years. Several exceptions from the early half of the 2010s exist, though. **Delort et al (2011),** for instance, described a novel classification technique they create to train a classifier from a partially labeled corpus to moderate inappropriate content on online discussion sites, implement it, and test it against baseline techniques.

Berger and Strathearn (2013) designed a scoring system to measure "influence" and "exposure" based on Twitter interactions, which they tested using 3,542 Twitter accounts which followed 12 known pro-IS accounts. They found that high scores of influence and exposure showed a strong correlation to engagement with the extremist ideology. **Berger and Morgan**

(2015) similarly sought a demographic snapshot of IS supporters on Twitter, as well as to outline a methodology for detecting pro-IS accounts. They trained a *machine learning (ML) classifier* on Twitter accounts, finding that pro-IS supporters could be identified from their profile descriptions. More in the area of AI-led analysis as opposed to detection, **Lara-Cabrera et al (2017)** employed a *lexicon-based approach*, tested on posts from pro-IS users, in an effort to translate a set of indicators found in social science models into a set of computational features to identify the characteristics of users at risk of radicalization. They find these indicators (frustration, negative content, perception of discrimination, negative ideas of Western society, and positive ideas about Jihadism) do indeed characterize radicalize users and promote their use as features to create *ML classifiers* for the automatic classification of users at risk of radicalization. **Ashcroft et al (2015)** trained *SVM*, *Naïve Bayes*, and *Adaboost* classifiers based on stylometric (n-grams, hashtags, word frequency, etc.), time-based, and sentiment features in an effort to automatically detect messages released by jihadist groups on Twitter. They found that Fridays are a key date to spread radical tweets and, more broadly, that automatic detection is viable but cannot replace human analysts.

Saif (2017) sought to create classifiers able to automatically identify pro-IS users in social media, finding that classifiers trained on *semantic features* outperformed those trained from lexical, sentiment, topic, and network features. **Fernandez and Alani (2018)** explore the use of semantic context to create more accurate radicalization detection methods, using semantic extraction of entities, entity types, topics, and categories from a knowledge graph to model context, and incorporating such context as features into *SVM*, *Naïve Bayes*, and *Decision Tree classifiers.* They find that semantic information can help to better understand the contextual variances in which radicalization terms are used when conveying 'radicalized meaning' vs. when not. **Fernandez et al (2018)** use *word vectors* to model the micro, meso, and macro radicalization influence in an effort to measure the influence of online radicalization that a user is exposed to and design a computational method based on the social science theory of roots of radicalization, finding that technological solutions to track radicalization need to leverage theoretical models more closely.

A few scholars have proposed approaches that focus on the *prediction* of online radicalization. **Magdy et al (2016)** trained an *SVM* classifier based on bag-of-words features, including individual terms, hashtags, and user mentions, tested on a dataset of Twitter users who authored or shared tweets mentioning IS, hoping the approach could predict future support or opposition to IS. They find that indeed, pro- and anti-IS users can be identified before they voice explicit support or opposition. **Ferrara et al (2016)** propose a computation framework for detection and prediction of adoption of radical content and interaction with pro-IS accounts. They use *random forest* and *logistic regression classifiers* and prediction based on user metadata and activity features, time features, and features based on network statistics. They find that the ratio of retweets to tweets, the average number of hashtags adopted, the sheer number of tweets, and the average number of retweets generated by each user rank very high in predictive power.

Nouh et al (2019) identify textual, psychological, and behavioral signals which together allow for the classification of radical messages. They analyze propaganda material published by extremist groups and create a contextual text-based model of radical content; build a model of psychological properties inferred from these materials; and evaluate the models on Twitter to determine the extent to which it is possible to automatically identify online radical tweets. They find that radical users do exhibit distinguishable properties; that psychological properties are among the most distinguishing; and that textual models using vector embedding features significantly improve detection over term frequency-inverse document frequency (TF-IDF) features.

The quantity of publications in computer science concerning methods of automated analysis, detection, prediction, flagging, and removal of content has ballooned dramatically with each passing year, particularly in 2020 and 2021. **Pandey et al (2021)** present a novel on-device solution for detecting NSFW images, from conventional pornography to semi-nude imagery. By curating a dataset of namely nude, semi-nude, and safe images, the authors create an ensemble of object detector and classifier with 95% precision and 0.2% false positive rate. Similarly, **Singh et al (2019)** address the tactic employed by malicious uploaders of placing violent or sexually explicit scenes within a video to catch children's attention by leveraging a *Long Short-Term Sectionry* (LSTM) based autoencoder to learn effective video representations of video descriptors obtained from *VGG16 Convolutional Neural Network* (CNN); they find that deep learning (DL) approaches are significantly more effective at combatting such content. **Hilscher et al (2021)** present a *service-based approach* towards content moderation of digital visual media while browsing webpages which enables the automatic analysis and classification of possibly offensive content (e.g. images of violence, nudity, or surgery) and applies common image abstraction techniques at different levels of abstraction to these to lower their affective impact. This system is implemented using a microservice architecture accessible via a browser extension installable in most modern web browsers, enabling greater parental control.

Ali et al (2020) propose a novel automated racial microaggression detection method using machine-learning algorithms. Given that microaggressions and hate-speech use very different features to convey their sentiment, the authors argue that conventional hate speech detection methods do not suffice. They employ several conventional ML algorithms used for text classification and sentiment analysis, ultimately finding that microaggressions can be detected from text (and perhaps then sent for human review), and that it may not be viable to use currently available hate-speech data for this type of detection.

Rahaman et al (2021) recent article on sarcasm detection in Tweets is another example of a fairly narrow, niche exploration of the potential of AI in content moderation. Though seemingly trivial, sarcasm poses a unique challenge insofar as it is difficult to detect due to the deliberate ambiguity of words (even for humans). Existing approaches to automatic sarcasm detection primarily rely on lexical and linguistic cues, but have produced minimal improvement in terms of accuracy, leading the authors to propose a new, robust system. They use four sets of features and various types of sarcasm commonly used in social media to train an ML model, ultimately showing that Decision Tree and Random Forest outperform other supervised machine learning algorithms in terms of accuracy.

Halfaker and Geiger (2020) developed ORES, an algorithmic scoring service that supports real-time scoring of Wikipedia edits using multiple independent classifiers trained on different datasets, which would decouple activities typically performed by engineers (e.g. choosing or curating training data, building models to serve predictions, auditing predictions, and developing interfaces or automated agents that act on those predictions. **Shelnutt (2021)** explores the effectiveness of *emotional analysis* as a means to automatically moderate content or flag content for manual moderation. He indeed finds that the moderation tools developed for the paper reduce the amount of posts that require human review by preemptively moderating posts using an emotional analysis algorithm in conjunction with a moderator-supplied configuration that defines tolerances of each emotion.

Another concern in the realm of content moderation is copyright infringement, which accordingly has been explored by computer scientists and scholars seeking to improve present systems. **Gray and Suzor (2020)** employ a massive dataset of YouTube videos to explore how digital and computational methods can be improved and leveraged in large-scale moderation systems. Using the *BERT* language model to trained an *ML classifier* to identify videos in categories that reflect ongoing controversies in copyright takedowns, they find that a simple neural network infrastructure, a pre-trained BERT language model, and substantial cloud processing power are able to achieve satisfactory performance on a multiclass classifier over short texts with a relatively small number of training examples.

In a rare work concerning content moderation outside the realm of platforms or malicious content, **Vivod (2020)** develops and illustrates a machine learning-based approach to message classification that could be used in school children's' book discussions, ultimately striving to predict whether a message posted in the discussion is relevant, and indeed finding it is feasible.

In conclusion, it is worth emphasizing the difficulty scholars face in studying online content moderation stemming from the lack of transparency provided by the platforms themselves. Though in the last few pages I have outlined potential future tools proposed by scholars to better moderate harmful online content, without a comprehensive understanding of which tools platforms do presently use, it's difficult to discern what precisely constitutes an "advancement." It is not unreasonable to think massive platforms like Facebook consult the popular computer science literature in search of innovations, nor to think such platforms already have implemented propriety automated detection systems more advanced than we are presently aware. Accordingly, the following section will conduct an individualized literature review of the major online content platforms and what we *do* know about how they moderate content, in hopes to provide the present section on advancements in the field some necessary contextual perspective.

Bibliography

Ali, Omar, Scheidt, Nancy, Gegov, Alexander, Haig, Ella, Adda, Mo, and Aziz, Benjamin. 2020. "Automated Detection of Racial Microaggressions Using Machine Learning." *2020 IEEE Symposium Series on Computation Intelligence (SSCI)*. 2477-2484. https://ieeexplore.ieee.org/abstract/document/9308569/authors#authors

Anderson, Jessica, Stender, Matthew, Myers West, Sarah, and York, Jullian C. 2016.

"Unfriending Censorship: Insights from Four Months of Crowdsourced Data on Social Media Censorship." *Onlinecensorship.org*. https://onlinecensorship.org/news-and-analysis/onlinecensorship-org-launches-first-report-download

Bickert, Monika and Fishman, Brian. 2018. "Hard Questions: How Effective Is Technology in Keeping Terrorists off Facebook?" *Facebook*. https://about.fb.com/news/2018/04/keeping-terrorists-off-facebook/

Chen, Thomas M. 2021. "Automated Content Classification in Social Media Platforms." In *Securing Social Networks in Cyberspace*, ed. Al-Sakib Khan Pathan. CRC Press. https://books.google.com/books?hl=en&lr=&id=wnY8EAAAQBAJ&oi=fnd&pg=PA53&dq=computer+science+content+moderation+machine&ots=ufl7ZdgQzy&sig=PsaRI7eg5S-8-55nwMig95U_oJg#v=onepage&q=computer%20science%20content%20moderation%20machine&f=false

Crawford, Kate and Gillespie, Tarleton. 2016. "What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media & Society* 18 (3): 410-428. https://journals.sagepub.com/doi/abs/10.1177/1461444814543163

Duarte, Natasha, Llanso, Emma, and Loup, Anna. 2017. "Mixed Message? The Limits of Automated Social Media Content Analysis." *Center for Democracy & Technology*. https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/

Fernandez, Miriam and Alani, Harith. 2018. "Contextual Semantics for Radicalisation Detection on Twitter." In *Semantic Web for Social Good Workshop (SW4SG) at International Semantic Web Conference.* CEUR. http://oro.open.ac.uk/56501/

Fernandez, Miriam, Asif, Moizzah, and Alani, Harith. 2018. "Understanding the Roots of Radicalisation on Twitter." *Proceedings of the 10th ACM Conference on Web Science*. ACM. https://dl.acm.org/doi/abs/10.1145/3201064.3201082

Fernandez, Miriam and Alani, Harith. 2021. "Artificial Intelligence and Online Extremism: Challenges and Opportunities." In *Predictive Policing and Artificial Intelligence*, eds. John McDaniel and Ken Pease. http://oro.open.ac.uk/69799/

Gallacher, John. 2020. "Automated Detection of Terrorist and Extremist Content." In *Extreme Digital Speech: Contexts, Responses and Solutions* eds. Bharath Ganesh and Jonathan Bright. *Vox-Pol*. https://www.voxpol.eu/download/vox-pol_publication/DCUJ770-VOX-Extreme-Digital-Speech.pdf

Gehl, Robert W., Moyer-Horner, Lucas, and Yeo, Sara K. 2016. "Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science." *Television & New Media* 18 (6): 529-547. https://journals.sagepub.com/doi/full/10.1177/1527476416680453?casa_token=ryU0oiMnNwwAAAAA%3A3TuX_3QBHy50P3zLLp0hGnIWEVPNbDstYUQMp1g2tVV1Bp_PbFzytP3yZtRMuO1rbKJGuJkpwxY

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the*

Hidden Decisions That Shape Social Media. New Haven: Yale University Press. https://www.degruyter.com/document/doi/10.12987/9780300235029/html

Gillespie, Tarleton. 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7 (2). https://journals.sagepub.com/doi/full/10.1177/2053951720943234

Gillespie, Tarleton, Aufderheide, Patricia, Carmi, Elinor, Gerrard, Ysabel, Gorwa, Robert, Matamoros-Fernández, Ariadna, Roberts, Sarah T., Sinnreich, Aram, and Myers West, Sarah. 2020. "Expanding the Debate about Content Moderation: Scholarly Research Agenda for the Coming Policy Debates." *Internet Policy Review* 9 (4). https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy

Gorwa, Robert, Binns, Reuben, and Katzenbach, Christian. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): 1-15. https://journals.sagepub.com/doi/full/10.1177/2053951719897945

Gray, Joanne E. and Suzor, Nicolas P. 2020. "Playing with Machines: Using Machine Learning to Understand Automated Copyright Enforcement at Scale." *Big Data & Society* 7 (1). https://journals.sagepub.com/doi/full/10.1177/2053951720919963

Halfaker, Aaron and Geiger, R. Stuart. 2020. "ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia." *Proceedings of the ACM on Human Computer Interaction* 4 (CSCW2), no. 148: 1-37. https://dl.acm.org/doi/abs/10.1145/3415219

Hilscher, Moritz, Tjabben, Hendrik, Rätz, Hendrik, Semmo, Amir, Besançon, Lonni, Döllner, Jürgen, and Trapp, Matthias. 2021. "Service-based Analysis and Abstraction for Content Moderation of Digital Images." Submitted for presentation at *Graphics Interface 2021 Conference Second Cycle*. https://openreview.net/forum?id=4j3avB-mrk

Jhaver, Shagun. 2020. "Identifying Opportunities to Improve Content Moderation." PhD diss., Georgia Institute of Technology. . https://smartech.gatech.edu/handle/1853/62779

Keller, Daphne and Leerssen, Paddy. 2020. "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation." In *Social Media and Democracy: The State of the Field, Prospects for Reform*, eds. Nathaniel Persily and Joshua A. Tucker, 220-248. Cambridge: Cambridge University Press. https://books.google.com/books?hl=en&lr=&id=TgH3DwAAQBAJ&oi=fnd&pg=PA220&dq=content+moderation+removal&ots=3BHp00akHt&sig=G6JcJVPOuSw5fEQSTuJ4u0i-9hc#v=onepage&q=content%20moderation%20removal&f=false

Mashechkin, I.V., Petrovskiy, M.I., Tsarev, D.V., and Chikunov, M.N. 2019. "Machine Learning Methods for Detecting and Monitoring Extremist Information on the Internet." *Programming and Computer Software* 45. https://link.springer.com/article/10.1134/S0361768819030058

Matias, J. Nathan. 2019. "The Civic Labor of Volunteer Moderators Online." *Social Media + Society* 5 (2). https://journals.sagepub.com/doi/full/10.1177/2056305119836778

Nouh, Mariam, Nurse, Jason R.C., and Goldsmith, Michael. 2019. "Understanding the Radical

Mind: Identifying Signals to Detect Extremist Content on Twitter." *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. 98-103. https://ieeexplore.ieee.org/abstract/document/8823548

Pandey, Anchal, Moharana, Sukumar, Mohanty, Debi Pasanna, Panwar, Archit, Agarwal, Dewang, Thota, Siva Prasad. 2021. "On-Device Content Moderation." *Cornell University: arXiv.org*. https://arxiv.org/abs/2107.11845

Postigo, Hector. 2009. "America Online Volunteers: Lessons from an Early Co-Production Community." *International Journal of Cultural Studies* 12 (5): 451-469. https://journals.sagepub.com/doi/abs/10.1177/1367877909337858

Rahaman, Arifur, Kuri, Ratnadip, Islam, Syful, Hossain, Md. Javed, and Kabir, Mohammad Humayun. 2021. "Sarcasm Detection in Tweets: A Feature-based Approach using Supervised Machine Learning Models." *International Journal of Advanced Computer Science and Applications (IJACSA)* 12 (6): 454-460. https://www.researchgate.net/profile/Mohammed-Kabir/publication/353078952_Sarcasm_Detection_in_Tweets_A_Feature-based_Approach_using_Supervised_Machine_Learning_Models/links/60ed96ff9541032c6d370b4c/Sarcasm-Detection-in-Tweets-A-Feature-based-Approach-using-Supervised-Machine-Learning-Models.pdf.

Risch, Julian and Krestel, Ralf. 2020. "Toxic Comment Detection in Online Discussions." In *Deep Learning-Based Approaches for Sentiment Analysis*. Algorithms for Intelligent Systems. Singapore: Springer. https://link.springer.com/chapter/10.1007/978-981-15-1216-2_4

Roberts, Sarah T. 2014. "Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation." PhD diss. University of Illinois at Urbana-Champaign. https://www.proquest.com/openview/eca12aca40888497c21007072a1a7322/1?pq-origsite=gscholar&cbl=18750

Roberts, Sarah T. 2016. "Commercial Content Moderation: Digital Laborers' Dirty Work." In *The Intersectional Internet: Race, Sex, Class and Culture Online*, eds. S.U. Noble and B. Tynes. Peter Lang Publishing. https://ir.lib.uwo.ca/commpub/12/?utm_source

Roberts, Sarah T. 2018. "Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation." *First Monday* 23 (3). https://journals.uic.edu/ojs/index.php/fm/article/view/8283

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press. https://www.degruyter.com/document/doi/10.12987/9780300245318/html

Shelnutt, John. 2021. "Applying Emotional Analysis for Automated Content Moderation." B.S. thesis, *University of Arkansas*. https://scholarworks.uark.edu/csceuht/93/

Singh, Spandana. 2019. "Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content." *New America: Open Technology Institute. https://www.newamerica.org/oti/reports/everything-moderation-*

analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

Singh, Shubham, Kaushal, Rishabh, Buduru, Arun Balaji, and Kumaraguru, Ponnurangam. 2019. "KidsGUARD: Fine Grained Approach for Child Unsafe Video Representation and Detection." *Presented at* SAC '19: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. *Association for Computing Machinery*, New York, NY. https://dl.acm.org/doi/abs/10.1145/3297280.3297487

Srinu, B., Swetha, M., Pranav, J.D., and Reddy, K.M. 2020. "Classification for Detecting Insulting and Abusive Content." *The International Journal of Analytical and Experimental Modal Analysis* 12 (4): 1692-1697. https://www.researchgate.net/profile/Dheeraj-Pranav-2/publication/341398253_Classification_for_Detecting_Insulting_and_Abusive_Content_The_International_journal_of_analytical_and_experimental_modal_analysis/links/5ebe3c67a6fdcc90d6756d91/Classification-for-Detecting-Insulting-and-Abusive-Content-The-International-journal-of-analytical-and-experimental-modal-analysis.pdf.

Udupa, Sahana, Hickok, Elonnai, Maronikolakis, Antonis, Schuetze, Hinrich, Csuka, Laura, Wisiorek, Axel, and Nann, Leah. 2021. "Artificial Intelligence, Extreme Speech, and the Challenges of Online Content Moderation." AI4Dignity Project. https://doi.org/10.5282/ubm/epub.76087

Vivod, Jernej. 2020. "Using Machine Learning and Natural Language Processing Techniques to Analyze and Support Moderation of Student Book Discussions." *Cornell University: arXiv.org*. https://arxiv.org/abs/2011.11712

Zhu, Wanzheng, Gong, Hongyu, Bansal, Rohan, Weinberg, Zachary, Christin, Nicolas, Fanti, Giulia, and Bhat, Suma. 2021. "Self-Supervised Euphemism Detection and Identification for Content Moderation." *Presented at 42nd IEEE Symposium on Security & Privacy*. https://arxiv.org/abs/2103.16808

9. Differences between Platforms

The following section provides a more opinionated overview of the literature.

Prepared primarily by LB

Background

The prior sections distinguished between the major social media and user content platforms (e.g. Facebook/Instagram, YouTube, and Twitter) who maintain robust automated and human-led moderation systems—and smaller platforms which have historically relied on volunteer moderation and/or moderated content less aggressively (e.g. Reddit, Discord, Gab, Parler, Telegram). Nonetheless, the major platforms are not a monolith, and though they have thus far been far less transparent with respect to their moderation systems than scholars would like, noteworthy differences between them exist. This is partly due to differences in leadership, but also because each platform has different volumes and types of content which often require a unique approach. They are also ever evolving; whereas video content moderation once made up a small subsection of Facebook's moderation infrastructure, they now receive over 4B video views every day. Or, to offer another example, consider that copyright enforcement presumably makes up a much larger portion of the moderation work at YouTube than it does for other companies.

Accordingly, this section briefly outlines some available scholarship on how the major companies go about moderating content uploaded to their platforms, and how those policies differ between them, as an addendum to the more extensive sections written over the last month. As previously mentioned, research into this area is inherently limited by transparency. Furthermore, though I did spend some additional time working on this section outside of my hours, it is incomplete, as my work has come to a close. Accordingly, this section may not be as comprehensive an overview of contemporary literature as the others, particularly sections 2-4. Nevertheless, it provides a solid baseline built upon the sources I have been able to gather, within which I imagine more sources of value may be identified.[1]

[1] In particular, see: **Singh (2019)**; **Gorwa, Binns, and Katzenbach (2020)**; and **Chen (2021)**.

Facebook

Of the platforms studied in the present section (as well as of those not included), Facebook has by far the largest content moderation operation. Community Standards on Facebook are grouped into five broad categories **(Chen 2021)**:

I. **Violence and criminal behavior** (e.g. incitement to violence; services for hire to kill or kidnap; threats to harm; instructions for weapons; violence due to democratic processes' propaganda for terrorism, hate, etc.; harm against animals; vandalism and hacking; sale of regulated goods; fraud and deception; stolen exams or answer sheets; fake reviews; scams; etc.)

II. **Safety** (e.g. instructions for suicide or self-harm; child exploitation; child abuse; sexual exploitation of adults; cyberbullying and harassment; privacy violations; etc.).

III. **Objectionable content** (e.g. hate speech; incitement to violence; graphic imagery (with certain exceptions, which are labeled and age restricted); nudity and sexual imagery; sexual solicitation; etc.).

IV. **Integrity and authenticity** (e.g. impersonation; spam; hacking; sharing or misusing accounts; fake news; media manipulated to mislead; etc.).

V. **Intellectual property** (e.g. copyright violation).

Facebook distinguishes between public figures and private individuals; the former are given less protection against critical commentary in the name of encouraging public discussion. Private individuals are given more protection, such as posts "meant to degrade or shame," and minors are given the most protection.

According to **Gorwa, Binns, and Katzenbach (2020)**, who recently wrote on the automation of content moderation, Facebook employs the following AI/computer programming technologies for the respective issue area. It is important to note that this information is gleaned from public reporting, and thus these platforms may not be report proprietary tools or have developed more sophisticated ones since.

A. **Terrorism**
   a. Shared Industry Hash Database (SIHD)
   b. ISIS/Al-Qaeda Classifier
B. **Violence**
   a. Community standards classifiers
C. **Toxic speech**
   a. Community standards classifiers
D. **Copyright**

  a. Rights manager

E. **Child abuse**
  a. PhotoDNA

F. **Sexual content**
  a. Non-consensual intimate image classifier
  b. Nudity detection

G. **Spam & automated accounts**
  a. Immune system

In that work, the authors breakdown a notable, unique algorithmic moderation system employed by each platform; for Facebook, they choose their toxic speech classifier (under community standards classifier). To deal with hate speech and bullying concerns, Facebook leveraged natural language processing (NLP) into a predictive classifier targeting text-based content. Of the notable systems unpacked by authors, it is the only one which uses deep-learning techniques. Humans still play a role in this moderation system—labeling the training data and setting the parameters for the predictive model, and making the takedown decision based on flags. In other words, Facebook predominately uses their toxic speech classifier as an augmentation of human moderation, not a replacement. Still, the platform has found that pattern detection is most effective for images rather than text, which can more easily be manipulated in order to evade detection and which requires greater contextual understanding to evaluate.

In September 2019, Facebook reported that the use of machine learning classifiers "led to the removal of more than 26 million pieces of content related global terrorist groups like ISIS and al-Qaeda in the last two years, 99% of which we proactively identified and removed before anyone reported it." Encouraging as this is, **Gillespie (2020)** argues that such statistics touted by platforms are misleading, as they imply that machine learning techniques are accurately spotting new instances of abhorrent content, whereas, at least presently, the overwhelming majority of what is automatically identified are copies of content that has already been reviewed by a human moderator.

It is important to emphasize that Facebook first deploys automated tools during the *ex-ante* stage of content moderation; when a user submits content to the site (e.g. a photograph), it is immediately screened in an automated process which uses digital hashes (outlined in detail in the previous two sections) to proactively identify and block content that matches existing hash databases for content such as child sex abuse and terrorism-related imagery. Once the content has been posted, the company engages in *ex-post* proactive moderation as it employs a different set of algorithms to screen and identify objectionable content. Such algorithms analyze content to identify similarities and patters in images, words, and behaviors associated with objectionable content.

If the system identifies that the content violates Community Standards, it may be automatically removed; however, if the algorithm is unsure, it will be routed to a human moderator. If a user flags content before the company is able to identify it, this flag also informs the ML models. Facebook's ML tools create a predictive score which, dependent upon it, will be flagged, with higher scores placing them higher in the queue for priority review. To audit the accuracy of automated decision-making, Facebook calculates two primary metrics: precision and recall. Precision measures the percentage of posts that were correctly labeled as violations out of all the posts that were *labeled* as violations, and recall measures the percentage of posts that were correctly labeled as violations out of all the posts that were *actually* violations. Nonetheless, there is little publicly available/known about Facebook's specific classifiers, the accuracy of their enforcement system, and the error and reversal rates, substantially inhibiting the amount of insight which can be gleaned from this audit process **(Singh 2019)**.

Twitter

The Twitter Rules are grouped into four categories (Chen 2021; Twitter 2021):

I. **Safety** (e.g. incitement to violence; child exploitation; abuse and harassment; hate speech; promotion of suicide or self-harm; graphic or violent content; illegal goods or services; etc.).

II. **Privacy** (e.g. stolen information; non-consensual nudity).

III. **Authenticity** (e.g. platform manipulation; spam; manipulating or interfering in democratic processes; impersonation; manipulated media).

IV. **Third-party advertising in video content**. This clause is used to protect tweets by public officials (notably President Trump) which offers a rationale: "we recognize that sometimes it may be in the public interest to allow people to view tweets that would otherwise be taken down. We consider content to be in the public interest if it directly contributes to understanding or discussion of a matter of public concern. At present, we limit exceptions to one critical type of public interest content- tweets from elected and government officials."

Gorwa, Binns, and Katzenbach (2020) outlined the following about Twitter's automated moderation tactics:

A. **Terrorism**
   a. Shared Industry Hash Database (SIHD)
B. **Toxic speech**
   a. Quality filter
C. **Child abuse**
   a. PhotoDNA
D. **Sexual content**
   a. Sexual content interstitial
E. **Spam & automated accounts**
   a. Proactive tweet and account detection
   b. Quality filter

For Twitter's unique automated approach, the authors unpack the quality filter used to deal with spam and harassment. Like Facebook's toxic speech classifier, the Twitter quality filter uses natural language processing (NLP) for prediction and requires humans to label training data and set the parameters for the predictive model. It does not, however, require humans make takedown decisions based on flags, and it also targets account information in addition to text.

Twitter developed the quality filter following years of criticism for insufficiently combatting harassment, which tries to predict whether content may be low-quality, spammy, or automated. Twitter is generally more hesitant to moderate than some of its counterparts, which Gorwa, Binns, and Katzenbach (2020) attribute to their First Amendment stance on freedom of expression; accordingly, the quality control filter was designed not to remove content, but rather to render it less visible.

YouTube[2]

Community standards on YouTube are grouped into four categories (Chen 2021; YouTube 2021):

I. **Spam, misleading practices, and scams** (e.g. video spam; comments spam; misleading metadata; incitement to interfere with democratic processes; hacked materials; impersonation; links to websites with banned content).

[2] For a summary of scholarship on YouTube's scraper—as well as the scrapers used by other platforms—see Sections 2 and 3.

II. **Sensitive content** (e.g. nudity and sexual content (art and education excepted); violent imagery; misleading thumbnails; harmful or stressful acts involving minors; cyberbullying of minors; promotion of self-harm).

III. **Violent or graphic content** (e.g. extremist content; dangerous challenges or pranks; incitement to violence or harm; illicit drugs; eating disorders; incitement to theft or hacking; dangerous remedies; bypassing payment for services; graphic depiction of bodily functions or corpses to shock or disgust; hate speech; cyberbulling and harassment).

IV. **Regulated content** (e.g. sale of illegal goods or services; counterfeits; illicit drugs; explosives; endangered animals; weapons; nicotine; sex services; instructions for firearms and weapons."

YouTube employs the following tools for the respective challenges:

A. **Terrorism**
   a. Shared Industry Hash Database (SIHD)
   b. Community Guidelines (CG) ML classifiers

B. **Violence**
   a. CG ML classifiers

C. **Toxic speech**
   a. CG ML classifiers

D. **Copyright**
   a. Content ID

E. **Child abuse**
   a. Content safety API
   b. PhotoDNA

F. **Sexual content**
   a. CG ML classifiers

G. **Spam & automated accounts**
   a. CG ML classifiers

The authors look particularly at YouTube's unique Content ID system, which has used hash-matching techniques (outlined in detail in sections 3 and 4) to target both audio and video content in order to prevent copyright infringement on the platform for more than a decade. They developed this tools in face of a growing threat from legal challenges, and it is highly tilted towards the preferences of the copyright holders it intends to pacify. After uploading audio or video content, copyright holders have the ability to select whether they wish to take down or receive a portion of the advertising revenue from content that matches the system's hashes. Such individuals have little recourse, and reversing those decisions is extremely difficult.

Much like its peers, YouTube remains secretive about the specific technological implementation of proprietary automated moderation systems, but some characteristics can be gleaned from public available material. Content ID is unique in that it allows copyright holders to upload the material that will be a) searched against existing content on YouTube and b) added to a hash database and used to detect new uploads of that content. Such systems are able to identify not only multiple uploads of, for instance, the same music video, but also recordings of live performances of that song. Through perceptual hashing, the resulting fingerprints reflect characteristics of audio or video content down to particular notes, frequency values, amplitudes, volumes, etc. Though concerns are prevalent regarding the ethics of Content ID, and whether it misidentifies instances of fair use, the tools has overall been sufficiently successful that Facebook (in 2016) and Instagram (in 2018) followed suit with their deployment of the Rights Management platform, which features similar functionality to Content ID (Singh 2019).

Instagram

Instagram (owned by Facebook), offers the following community guidelines (Instagram 2021):

I. **Share only photos and videos that you've taken or have the right to share.**

II. **Post photos and videos that are appropriate for a diverse audience.** Nudity is not allowed, except in the context of breastfeeding, birth giving, and after-birth moments, health-related situations, an act of protest, or in paintings or sculptures.

III. **Foster meaningful and genuine interactions**. Do not artificially collect likes, followers, or shares, post repetitive comments or content, or repeatedly contact people without their consent. Do not offer anything in exchange for likes or other engagement.

IV. **Follow the law.** No sales of sexual services, firearms, live animals, alcohol, drugs, and tobacco productive are allowed. Promotion of gambling, sexual content involving minors, or revenge porn are not allowed.

V. **Respect other members of the Instagram community**. No targeting, encouragement of violence, hate speech, threats, repeated unwanted messages, etc.

VI. **Maintain our supportive environment by not glorifying self-injury.** Content supporting ED behavior and other self-injury is not allowed.

VII. **Be thoughtful when posting newsworthy events**. If sharing newsworthy events which involve graphic images, captioning with a graphic violence warning is recommended; the sharing of graphic images for sadistic pleasure or to glorify violence is never allowed.

Instagram employs the following tools for the respective challenges:

A. **Toxic speech**
  a. Comment filter

B. **Copyright**
  a. Rights manager

C. **Child abuse**
  a. PhotoDNA

D. **Spam & automated accounts**
  a. Comment filter
  b. False account detection

One method platforms employ to moderate content is through the regulation, banning, or censoring of particular *hashtags*; this strategy is most closely associated with Instagram, on which hashtags are the predominant method of searching for content from users that a person is not already following. Instagram moderates hashtags in four different ways: 1) a 'hard' block approach, in which a user visiting these tags will encounter an HTTP 404 error; 2) only displaying a selection of 'top posts' of a hashtag; 3) displaying an interstitial 'content warning' for tags related to eating disorders, self-harm, firearms, animal abuse, and at times providing links to third-party helplines; and 4) a 'silent' block approach, in which the platform presents a false statement that there are "no posts yet" on the tag.

Unlike YouTube and Twitter, Instagram provides no public application programming interface (API) that can be used to build a random sample of posts nor does it provide any public information about whether it has moderated a post, and there is little easily discernible pattern that explains why it chooses some forms of blocking over others. Moreover, the status of hashtags sometimes changes over time **Suzor (2020)**. This renders researching hashtag moderation by Instagram even more difficult than researching YouTube's scraper or Twitter's tester, but Suzor (2020) and The Observatory nonetheless track it, as well as how Instagram moderates posts.[3]

[3] In order to track which hashtags are blocked on Instagram, Suzor (and The Observatory) first developed a super-set of hashtags that had been reported as blocked by news outlets and

Though the present section provides some necessary delineation with respect to the four major social media platforms moderate content—especially algorithmically—it is not exhaustive. Then again, with the degree of transparency platforms offer, one cannot be truly exhaustive in answering this question. That being said, a more thorough review of the literature—including and especially platforms' own press releases and publications—might shed light on tools not mentioned in the present section, tools recently developed, and tools the existence of which was not yet known. Moreover, there are other platforms which moderate content, employing propriety tools worthy of analysis. In addition to the unique tools Gorwa, Binns, and Katzenbach (2020) look at summarized above, for instance, the authors look at the Perspective API system used by Google Jigsaw and the PhotoDNA system pioneered by Microsoft and now employed by all platforms mentioned.

Bibliography

Chen, Thomas M. 2021. "Automated Content Classification in Social Media Platforms." In *Securing Social Networks in Cyberspace*, ed. Al-Sakib Khan Pathan. CRC Press. https://books.google.com/books?hl=en&lr=&id=wnY8EAAAQBAJ&oi=fnd&pg=PA53&dq=computer+science+content+moderation+machine&ots=ufl7ZdgQzy&sig=PsaRI7eg5S-8-55nwMig95U_oJg#v=onepage&q=computer%20science%20content%20moderation%20machine&f=false

Facebook. 2021. "Community Standards." https://www.facebook.com/communitystandards/?ref=u2u (August 30, 2021).

Gillespie, Tarleton. 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7 (2). https://journals.sagepub.com/doi/full/10.1177/2053951720943234

Gorwa, Robert, Binns, Reuben, and Katzenbach, Christian. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): 1-15. https://journals.sagepub.com/doi/full/10.1177/2053951719897945

supplemented it with search results on a major search engine for the error messages that Instagram displays when it blocks a hashtag or hides it behind an interstitial content warning. They then check these 1,449 hashtags daily for availability and separately test 5000 hashtags that have been most frequently used with posts that use previously-blocked hashtags in the sample. This provided ~12K tags tested each day, and an eventual sample of 12.6K hashtags that have been blocked at least once.

Instagram. 2021. “Community Guidelines.” https://help.instagram.com/477434105621119 (August 30, 2021).

Singh, Spandana. 2019. “Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content.” *New America: Open Technology Institute. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/*

Twitter. 2021. “The Twitter Rules.” https://help.twitter.com/en/rules-and-policies/twitter-rules (August 30, 2021).