International Journal of Communication 19(2025), 998–1018 1932–8036/20250005

Creating a Cost to Spread Misinformation on Social Media

DREW B. MARGOLIN[1]
YUNYUN S. WANG
Cornell University, USA

This study analyzes systemic incentives that encourage the spread of misinformation on social media. We show how social media's algorithmic opacity and "open entry" nature incentivize users we call *unaccountable spread seekers* to overproduce misleading messages. Specifically, these users are encouraged to "propagation hunt," producing many messages to identify the few that go viral. We propose a remedy to this challenge called the attestation framework, in which users must explicitly declare their intent that a message be eligible to spread. By attesting, users agree to take on some liability for the harm this spread may cause. We show that attestation should substantially curb the activities of unaccountable spread seekers while having minimal impact on other users.

Keywords: misinformation, policy, platforms, incentives, attestation

Addressing the spread of misinformation and other harmful content on social media is an important and challenging public priority. However, proposals to date focus mostly on addressing messages that have already been produced: How can misinformation be detected? How should it be penalized? Who should be empowered, and using what standards, to moderate it? By contrast, less attention has been paid to the *incentive structure* that shapes the population of messages that are produced and injected into the system and on which further, more downstream measures act.

In this study, we show that incentives in the current system are contributing to the problem by encouraging certain users to produce a large quantity of potentially harmful messages, a practice we refer to as *propagation hunting.* To address this weakness, we identify a regulatory principle we call the *attestation framework* in which users must attest to the truthfulness of a post for it to be eligible to be spread widely via the platform. We argue that attestation should discourage propagation hunting, reducing the quantity of misinformation in the system, while minimally impacting good-faith users.

The argument is organized as follows: First, we identify the quantity problem––that platforms face too many messages and too little time to evaluate them. We show that these challenges are exacerbated

Drew B. Margolin: dm658@cornell.edu
Yunyun S. Wang: yw458@cornell.edu
Date submitted: 2023-10-27

[1] We wish to thank Dr. Tarleton Gillespie for his helpful feedback on this study.

by users we call *unaccountable spread seekers*, who are encouraged by the system to propagation hunt. We then briefly outline the limitations of common solutions to these problems before introducing the attestation framework. Attestation stipulates that if a user wishes his or her message to reach a large audience, they must bear some liability for the harm the message could cause. We show that attestation should discourage propagation hunting and more generally reduce the quantity of misinformation introduced onto platforms. We also show that attestation should have a limited impact on good-faith users. We do not argue that attestation is *the* solution to the problem of misinformation online, only that it is a viable and attractive remedy to the problems we identify.

The Quantity Problem

Hidden in plain sight in the discussion of how to regulate social media is the problem of quantity––how can any entity evaluate so much content? For example, in 2022, Twitter (now X) reported that there are 500 million tweets posted each day (*Twitter Usage Statistics—Internet Live Stats*, 2022). Facebook's 2.9 billion monthly active users post an even more daunting 350 million *images* per day (*Facebook MAU Worldwide 2022*, 2022).

The problem of regulating such a large quantity of behavior is a persistent challenge in the study of regulation (Feldman, 2018), and in fact, this quantity problem lies at the center of justification for the protections afforded to Internet companies under Section 230 of the Communication Decency Act (*Harvard Law Review*, 2018). Section 230 is based on the distinction between bookstores ("platforms") and publishers identified by the Supreme Court in *Smith v. California* (1959). In *Smith v. California*, the Court ruled that holding a bookstore owner liable for the content of each book they sold would create an enormous monitoring burden, as owners would need to have knowledge of every book in their stores to ensure they were not selling obscene material. The court noted that this burden would reduce public access to published material and information to limited numbers of books store owners personally knew weren't obscene (*Smith v. California*, 361 U.S. 147, 1959).

One way to limit the number of messages in need of evaluation would be for platforms to evaluate only those that spread broadly to the public. Oliver Wendell Holmes's famous statement that free speech should be limited when it presents a clear and present danger, such as "falsely shouting fire in a theater," argues that it is not the content of the speech that should be regulated but its potential for harm when it is presented to a large audience. Empirical research supports this intuition, as spreading messages to large audiences can have potentially more dangerous effects through such processes as emotional contagion (Coviello et al., 2014), reorienting collective attention (Shteynberg, 2015), and herding (Bikhchandani, Hirshleifer, & Welch, 1998).

However, choosing to evaluate only the messages that are likely to spread merely appears to solve the quantity problem. In the current social media ecosystem, no one knows *which messages* are going to reach virality. Few messages reach large audiences (Goel, Anderson, Hofman, & Watts, 2016). Although messages from individuals with large followings tend to have an advantage, this does not explain the bulk of viral content (Keller & Kleinen-von Königslöw, 2018). There are not a priori criteria for determining which messages will go viral (Bakshy, Hofman, Mason, & Watts, 2011), nor are there standard growth paths that

identify them (Goel et al., 2016). Rather, virality appears to depend on a property of a post that is only detectable after it is revealed (Cheng, Adamic, Kleinberg, & Leskovec, 2016). Users themselves show a limited ability to recapture virality in future posts (Guinaudeau, Munger, & Votta, 2022).

A related problem is that individual messages can be edited to convey the same ideas in new forms. As a message spreads, it is more likely to be adapted in this way (Leskovec, Backstrom, & Kleinberg, 2009), and thus a (harmful) message that reaches even a small audience has more chances at being edited into a more spreadable form. Thus, the level of spread at which a message becomes *potentially* harmful is substantially smaller than the level at which it is *actually* harmful. If companies must monitor potentially harmful messages to prevent harm, logically, this entails monitoring all messages.

An additional monitoring challenge is the infeasibility of evaluating potential harm early in a message's lifespan, before it spreads widely. As of this writing, the monitoring scheme for each of Meta (n.d.), TikTok (2022), and X (2023) deploys fact-checking by independent organizations and internal teams. This work, however, can demand hours or days of research, requiring steps such as contacting the claimant for corroborating evidence (Hassan et al., 2015). For example, during disasters, people often spread rumors immediately about specific dangers or culprits (Starbird, Maddock, Orand, Achterman, & Mason, 2014). Until these rumors are investigated, it is not possible to know with any certainty whether they are true or false. However, waiting longer to monitor messages is also a problem. The impact of exposure to misinformation is not entirely reversed by a subsequent exposure to a correction of that misinformation (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). Thus, by the time a post is proven false and labeled or removed, it may have already been seen by––and influenced––millions of viewers (Thorson, 2016).

In summary, platforms face a nearly insurmountable problem. All messages need to be checked as soon as possible before they spread. However, an earlier evaluation will result in a less accurate judgment.

Unaccountable Spread Seekers

The quantity problem is exacerbated by incentives for unaccountable spread seekers to produce excess messages. Unaccountable spread seekers are individuals, groups, or organizations who produce messages for the sake of spreading them widely but who are not governed by any source of accountability to keep them from producing social harm. Unaccountable spread-seeking has been a feature of political and commercial discourse for centuries (Varol & Uluturk, 2018). Examples may include foreign governments or insurgent groups who wish to sow discord or entities who profit from the spread of clickbait (Lawson & Kakkar, 2022).

As defined, the key characteristic of these users is the value they place on the virality of their messages above all else. They may value virality because the legitimacy of an idea grows as the number of people who have heard it increases (Hannan, Polos, & Carroll, 2007) or because repeated exposure to misinformation makes it more believable (Lewandowsky et al., 2012). Virality may help their causes through bandwagon effects, in which the belief that a candidate, party, or idea has a lot of support draws more

supporters to it (Kleinnijenhuis, van Hoof, & van Atteveldt, 2019), as well as entitativity effects, in which the size of a group supporting an idea makes it appear more formidable (Campbell, 1958).

However, unaccountable spread seekers are not typical or common as individuals. First, most people are not spread seekers. Research shows that, for most people, communication intensity generally is much lower for their larger sets of weak ties and is instead focused substantially within their own networks of limited size (Sutcliffe, Dunbar, Binder, & Arrow, 2012). Thus, for most people, there is a diminishing, rather than increasing, marginal benefit to reaching a larger audience.

Second, most spread-seeking individuals and organizations face some accountability for their actions or behaviors, an accountability that grows with their audience. For example, individuals may fear that if their messages reach large audiences, they may be judged according to a higher standard (Marwick & boyd, 2011). Similarly, many professionals and organizations are held to codes of conduct, such as journalistic ethics or medical ethics (De Bruin, 2016). In social media, however, individuals can gain widespread attention for their messages without anything "at stake," that is, any preexisting commitments to a social group or profession that they risk losing if they produce harmful content. We provide Table 1 to organize these conceptions and user groups. As shown in the table, unaccountable spread seekers are a particular type of user.

Table 1. User Groups.

	High Value From Spread	Low Value From Spread
High Accountability	Accountable spread seekers: e.g., journalists, medical professionals	Accountable nonspread seekers: e.g., Government officials, local organizations, small businesses
Low Accountability	Unaccountable spread seekers: e.g., fake journalists, propagandists, clickbait farms	Unaccountable nonspread seekers: e.g., Everyday users

Note. Users are categorized by the value they gain from spreading messages and their accountability for social impact.

The current media system encourages unaccountable spread seekers to produce an excess of messages, exacerbating the quantity problem. The first inducement to doing so is the low, almost frictionless cost of sending messages and creating accounts. For a user with an existing account, the marginal cost of producing a message is essentially zero. It is common, for example, for some users on X to produce hundreds of messages a day (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019). The barriers to entry as a spreader are also extremely low. As of this writing, the minimal credentials required to create an account on Facebook or Instagram (n.d.), TikTok (n.d.), and X (n.d.) is an e-mail address or phone number. User profile information can also be easily changed to avoid reputational consequences (Zannettou et al., 2019). False accounts with existing audiences can also be purchased (Busby, 2018). This "open entry" nature of the system means that getting access, at least in principle, to the ability to spread messages—the goal of the unaccountable spread seeker—is de minimis.

Spread seekers face the same uncertainty as platforms—they do not necessarily know which messages will spread. Unfortunately, "open entry" encourages a particularly problematic solution to this problem. Specifically, because it is difficult for spread-seeking users to know which messages will spread, and because the cost of producing messages is so low, it is in their interest to flood the system with messages in search of the most viral variant. We call this process of producing large numbers of messages to find the one viral "needle" in the large "haystack" of ignored messages *propagation hunting*. Empirical studies show this pattern. For example, an analysis of misinformation posts on (then) Twitter shows that a large portion of them are sent by a few accounts that send an enormous number each day (Grinberg et al., 2019). According to an estimate by Varol, Ferrara, Davis, Menczer, and Flammini (2017), 15% of Twitter accounts are bots. Lazer and colleagues (2018) also state that "Facebook estimated that as many as 60 million bots (7) may be infesting its platform" (p. 1095). These figures suggest that user accounts are created cheaply in an automated or semiautomated fashion.

Similarly, there is evidence of "fake news farms" that send variations on common themes to see which will be the most viral. A *Wired* report shows the extent of these operations, revealing over 100 pro-Trump websites filled with sensationalist fake news registered in a single town during his 2016 campaign (Subramanian, 2017). These operations often own hundreds of Facebook profiles purchased for reposting and sharing their own false content to drive up engagement and reach virality. As documented in the *Wired* story, experienced practitioners diligently craft their profiles and messages to produce viral content, what some have called the "firehose of falsehood" (Paul & Matthews, 2016, p. 1).

Limitations of Existing Approaches to Accountability

As stated above, unaccountable spread seekers are not new; rather, it is their potential to do harm that varies with the accountability system in place within a particular media regime. Here, we briefly review alternative systems that have been tried or are currently being implemented.

The Mass Media Regime

The previous media-discourse regime is typically referred to as "mass media" (O'Sullivan & Carr, 2017). The threat of unaccountable spread seekers was widely recognized after World War I, in which propaganda was rampant and gave rise to a cottage industry (Bernays, 2020). The mass media regime managed (suppressed) unaccountable spread seekers by creating barriers to entry. Only a limited few—those with the resources and capital to acquire means for mass communication—could spread their speech through these mass media channels. Most individuals could reach only a few audience members that were in their personal networks (Rainie & Wellman, 2012). Enormous audiences could still be reached, but only from very select sources—those with access to broadcast technology and infrastructure.

This few-to-many structure for reaching mass audiences is often described as a "gatekeeper" structure. As many who hailed the Internet and its many-to-many capabilities pointed out, the gatekeeper structure has an important limitation. It marginalizes voices that lack access to the spreading technology—individuals with fewer resources, social status because of race or gender, or politically marginal views. The

gatekeeper structure thus often created an allusion of deliberation and consensus based on substantial exclusion of alternative views (Fraser, 1990).

However, the exclusionary nature of the gatekeeper structure also disincentivized behaviors by would-be spread seekers because gatekeeper status was scarce and could be revoked. Gatekeepers had to establish themselves through mainstream institutions and adhere to collective ideas of social norms. For example, to spread misinformation, an individual would first have to establish credibility and achieve social standing to become a voice in mass media. Similarly, broadcast licenses could be suspended. The ability to spread could also be hampered by the withdrawal of social support from other gatekeepers (Burt, 2005). For example, the U.S. Post Office refused to deliver *Hustler* magazine (McTeague, 1988).

Thus, although access to the mass audience was disproportionately granted to the privileged, the scarcity of gatekeeping positions meant that it came with forms of explicit and implicit accountability that put downward pressure on the inclination to spread harmful messages; in essence, it limited the number of *un*accountable spread seekers. This is not to imply that the old regime was preferable, only to note that the otherwise laudable replacement of the gatekeeper structure has reinvited these harmful voices. While returning to the old structure is neither desirable nor feasible, the system would benefit from ways that reimpose constraints on would-be spreaders without once again excluding marginalized voices.

Current Regime—Flag and Remove

Current policies for addressing misinformation, such as those alluded to above at Meta, X, and TikTok, can be described as following a formula of "flag and remove." Messages are checked against third-party databases, flagged if the content is considered suspect, and then either downgraded for further distribution or removed entirely. For example, Meta writes:

> If you post content that one of our third-party fact-checkers rates False, Altered or Partly False, or we detect it as near identical, it may receive reduced distribution on Facebook, Instagram and Threads. . . . If you repeatedly post content rated False or Altered on Facebook or Instagram, we'll take several actions that will last for a subsequent 90 days. For example, we'll move all of your posts lower in Feed on Facebook or Feed and Stories on Instagram so people are less likely to see them. (Meta, n.d., p. 1)

Similarly, X states:

> For high-severity violations of the policy, including misleading media that have a serious risk of harm to individuals or communities, we will require you to remove this content. . . . In circumstances where we do not remove content which violates this policy, we may provide additional context on posts sharing the misleading media where they appear on X. (X, n.d.b, p. 1)

This "additional context" can include a "warning message," "reduce[d] visibility," and turning off "reposts." Additional consequences for repeat offenders can also include demonetization or account locks.

Although this system has useful elements, it remains directly vulnerable to propagation hunting by unaccountable spreaders. First, messages are not restricted or penalized until fact-checkers catch up with them. Thus, the greater the number of false messages produced, the smaller the proportion of them will be restricted. Second, the primary consequence is merely applied to the offending message itself (via deletion or distribution downgrading). But since messages have essentially zero production cost, the rational response from spread seekers is simply to create another message (or 100 messages) that might not get detected. Penalties at the user level are also weak. Repeat offenders may have their accounts restricted, but, just as it is easy to produce new messages, the cost of creating new accounts is also very small. Third, demonetization is only a disincentive to producers who are interested in earning money from their posts, but this is not the goal of unaccountable spread seekers. Their goal is spread itself. This system thus encourages the quantity problem.

Limits of Automated Evaluation

The quantity problem may appear less daunting given that tech companies can use automated filters guided by artificial intelligence. Since such processes are rapid and automatic, in principle they offer highly scalable ways to address the quantity problem. However, incentives for propagation hunting limit the likelihood that such strategies will be effective.

Algorithms are valued because of their consistency and reliability (Gillespie, 2014; Lazer, 2015), making them useful because their performance is predictable. However, this predictability enables "reverse engineering" (Frey, Albino, & Williams, 2018). If the algorithm is designed to flag messages with feature set A and let through messages with feature set B, then an individual with knowledge of this predictable response can use it to design messages that pass the test while serving his or her purpose. Thus, an effective monitoring algorithm must not only identify feature sets A and B that discriminate between "acceptable" and "harmful" information but also keep those feature sets *secret*.

Such nontransparency is problematic, however. In addition to giving companies unaccountable power and frustrating users (Patel & Hecht-Felella, 2021; Tobin, Varner, & Angwin, 2017), opacity actually encourages more propagation hunting. If platforms are going to remove some messages for unknown reasons, then spread seekers must produce more messages to find the variants that "sneak through." Given the almost limitless number of trials and errors afforded by open entry, *some* messages from spread seekers will inevitably pass the monitor's test. The opacity thus encourages an "arms race" between spread seekers and platforms (Yang et al., 2019), in which unaccountable spread seekers attempt to craft messages that reach virality by experimenting with different variants and observing what works best (Bakshy et al., 2011). This intensified propagation hunting produces a flood of messages and also increases transparency only for these users.

In summary, while automated detection can play an important role in a new system of accountability, it is not sufficient. Whereas the gatekeeper model grants the privilege of spreading to those

with wealth and power, the flag-and-remove-with-automated-detection model grants power to those with the aim and resources to produce message variants at scale.

Attestation

What is needed is a modification in the current incentive system, one that imposes a cost that will be borne by unaccountable spread seekers without restricting the freedom for most individuals to speak freely on platforms. We propose the attestation framework as one such remedy. In this section, we describe (1) how attestation works; (2) its expected impact on information flow; and (3) answers to common objections to the framework.

How Attestation Works

Definition and Rationale

In attestation, users explicitly declare whether they "attest" to each post they make. By attesting to a post, the user declares that he or she believes the post does not contain misinformation. This declaration then affords the post a privilege—it designates it eligible for spreading to the "public," that is, beyond the individual's personal network. Attested posts can be shared with others with one click ("resharing") and diffused to nonfollowers via algorithms ("trending") or hashtags, whereas unattested posts cannot. Unattested posts are shared with followers, but they are not searchable/discoverable to the general public.

The cost of attestation is responsibility for the content in the message. If an attested post is found to contain misinformation, the user is held responsible and may face penalties, including the need to compensate for damages assessed or, in extreme cases, criminal prosecution.

By default, messages are unattested. Users must take the affirmative step of attesting to a message, affirming its truth-value, taking explicit responsibility for it, and, in so doing, designating it eligible for spread.

A useful analogy for understanding attestation is the oath to be truthful taken by witnesses at trials. Attestation, like this oath, is an *assertatory oath*, a pledge about the veracity of a particular statement (De Bruin, 2016). Oath-taking is an important tool in the regulation of behavior, particularly when the cost of monitoring is high (Feldman, 2018). At the individual level, evidence suggests that oath-taking can encourage people to be more honest (Beck, Bühren, Frank, & Khachatryan, 2020), but that the effect is stronger when there is a penalty attached to violating the oath (Peer & Feldman, 2021).

Oaths are also useful because they can deter more professional or institutional bad-faith actors, whose behavior is not necessarily governed by the psychological responses of typical individuals. For example, the crime of perjury—the violation of the witness oath—emerged in response to concerns about a growing cottage industry of "maintenance," the paying of witnesses for testimony, an interesting parallel to the fake news farm of today (Gordon, 1980). One reason is that the logic of oath demands increased consequences for their violations. Assertatory oaths declare that a speaker's intentions for the speech are serious

(MacCormick, 1983), and that care has been taken by the speaker, reducing the need for audience skepticism, what Green (2001) calls *caveat auditor* ("listener beware"). Having made this promise, the violation is graver than mere deception (Green, 2001). It is a breach of an explicit trust and so warrants a greater punishment.

Implementation

The appropriate consequences—the exact costs of attesting to and thus spreading misinformation—would have to be worked out through legislation or other regulatory decisions, much as has been done for other kinds of serious deception, such as for deception by physicians (*Fraud & Abuse Laws*, 2021) or fraud in securities (*Securities Fraud Laws*, 2019).

For example, spreading misinformation during natural disasters (Starbird et al., 2014) could be more serious than spreading false gossip about politicians. As with other industries, increasing penalties, including eventual criminal charges and imprisonment, can then be used to discourage organized, corporate offenders with substantial financial resources for whom fines are a minimal deterrent.

That said, attestation does not introduce a radical change to platform governance. As cited above, most platform policies already use reduction of spread as a consequence when information is suspected of being false. WhatsApp has also preempted spread by with stricter forwarding limits following a series of deaths associated with rumors circulating on its platform (Chen, 2020). Platforms also recognize the concept of aligning privileges and responsibilities. For example, YouTube recently increased the benchmark for monetization, raising the previous threshold to join the YouTube Partner Program (Kain, 2018). This tiering establishes that there are privileges to this position that must be earned and can be taken away.

Nonetheless, attestation does require platforms to do three things that they do not currently do. The first is to collect sufficient information from attesters to make it possible to impose the penalties. Right now, platforms do not require such information, requiring only an e-mail address or phone number at account setup. Thus, the strictest penalty the platform can impose is internal—account locking/suspending, which is hardly a deterrent given the ease with which new accounts can be created. New rules would require platforms to collect enough information to make attestation rules enforceable. This data collection can have a deterrent effect on its own (Cecka, 2014); however, importantly, users do not need to provide any such details to hold an account or post to the platform. These details would be required only when they chose to attest—spread their speech to the public.

The second is to indicate the attestation status of each post. This will be done, at minimum, because users will not be able to engage in certain functions—such as resharing—on unattested posts. However, we recommend that the attestation status be made more prominent via a label or other signal, which indicates both the author's intent for (to spread or not to spread) and confidence in the message.

Third, the imposition of liability would require platforms be more transparent in their processes of detection. Current platform policies are vague and emphasize platform discretion. For example, as quoted above, Meta states that misinformation "may receive" a consequence; X states that they "may" take various contextual actions; and TikTok "may remove" videos found false by fact-checkers. This unilateral and opaque

discretion, while frustrating to users, is ultimately tolerable because the consequences pertain only to privileges the platform itself provides. But once penalties go beyond platform privileges, rules for what must be collected from attesters and what would constitute appropriate penalties would require government regulation proper, including safeguards and due process. As with oaths, by making consequences more serious, there is then greater responsibility.

Attestation does not change the terms of Section 230. Platforms are still not responsible for what is posted on them. They are simply required to provide *some* responsible party for all content that spreads on them, a distinction that Section 230 has let slip through the cracks. As such, platforms could also attest to content on their own, in essence "subsidizing" users to improve their experiences. For example, if a user does not attest to a post, but a platform wishes to promote it because of its high potential for engagement, the platform would have to assume the attesting role and the liability that go with it.

Thus, every message that spreads to the public has a responsible party attached—the party that intended for the message to spread to the public and thus takes responsibility for the consequences of that spread. Attestation accomplishes this aim without suppressing speech, only spread, and within the constraints and affordances of the current media regime's own affordances: the ability of users to label each individual post and platforms' ability to provide or deny spreadability on a post-to-post basis.

Impact on Information Flow

The attestation framework offers four benefits, specifically attestation should (a) reduce the number of inaccurate messages eligible to spread; (b) reduce the number of messages the platform must evaluate; (c) increase the accuracy of evaluations; and (d) enable platforms to share important, user-generated context for other users to consider when evaluating messages.

Proportion of Spreading Messages That Contain Misinformation

With attestation, users have an incentive to consult their private information about a post—did they invent it, is it a conjecture, or a provocation?—to filter what is made eligible for spread. Currently, the only reason to exercise restraint is reputational risks, which have force only for a subset of users (see Table 1). With attestation, even otherwise unaccountable users are encouraged to filter out irresponsible posts, reducing the number of false messages in the system as a whole. This may, for example, encourage many users to simply pay more attention to accuracy (Pennycook et al., 2021). Attestation should also suppress propagation hunting. By imposing a cost—the risk of being held liable for misinformation—on *each* message intended for spread, attestation reduces the attractiveness of the strategy of producing an overwhelming number of messages.

Number of Messages to Evaluate for Misinformation

Attestation should also reduce the number of messages platforms must scrutinize, as attested messages are only a subset of speech produced on a platform. With fewer messages to evaluate, there will be fewer absolute errors made, reducing overall harm. This limits the spread of misinformation from false

claims that go unidentified (Type II errors) and also reduces user frustration because of Type I errors made by platforms. The tradeoff for this benefit (fewer errors) is that fewer messages will spread overall. However, the messages that would no longer spread are specifically those that their own authors choose to withdraw from the "spreadable" pool.

Accuracy of Assessment of Message Information

Attestation should make the evaluation of individual messages more accurate by allowing the use of more post hoc information in judging the accuracy of messages. In the current system, messages are penalized (with reduced spread) only if they are already identified as problematic or suspicious based on what is known at the time. False claims that look true now will spread, even when the poster knows they are false.

Under attestation, by contrast, the threat of future liability creates an incentive that shapes the user's present posting decision. Just as a witness would fear consequences of perjury even if the perjury were not discovered immediately, the attesting user must consider the possibility that they will be found out and punished long after. This rationale is also similar to that used with promissory professional oaths (De Bruin, 2016), where individuals are given the autonomy to perform a task that is difficult to monitor and are held responsible if it is discovered later that it was done inappropriately (Zittrain, 2014). By allowing more time to gather relevant information, this post hoc remedy increases accuracy in identifying misbehavior.

This use of post hoc information also further discourages propagation hunting by delaying final judgment indefinitely. Currently, propagation hunters get immediate feedback on detection algorithms. With attestation, no such "final" decision is ever rendered. The propagation hunter knows only that his or her post has not *yet* been identified as harmful.

Additional Information for Interpreting Messages

Attestation also provides users with guidance for how to interpret posts. A label declaring the attestation status of the post provides viewers/readers with important contextual information. When an author chooses not to attest to a post, this signals that (a) they are not particularly confident in it and/or (b) they intend it for a more narrow, personalized audience (Marwick & boyd, 2011).

It should be noted that social media companies now apply several labels to content, such as that it is being fact-checked or may require context. However, the attestation label will stand out from these because it carries with it a different functionality, as unattested content will not be one-click shareable. Thus, it is reasonable to assume that users will notice this difference even if there is an overload of other labels.

Objections to Attestation

We do not wish to suggest that attestation is an exclusive answer to the problem of misinformation, only that it should be considered in part of an ensemble of solutions. That said, we think there is value in addressing three likely objections.

Objection 1—Injustice

One objection is that attestation is unjust (Wenzel, Okimoto, Feather, & Platow, 2008), in that it imposes costs on users even though it is the platforms that have benefited from tolerating the spread of misinformation. In response, we argue that, first, the focus of misinformation policy should be the reduction of future harm as a practical, forward-looking matter distinct from retrospective compensation for past harms. Also, by requiring that every post have a responsible party, attestation facilitates these remedies.

Second, by identifying a particular category of speech with a declared public interest—speech intended to spread—attestation introduces a legitimate regulatory interest without running afoul of Section 230. Legislative and regulatory attention is now needed to provide a fair process of adjudication over liability between these two private parties (companies, users). Most plainly, what should be the penalties for *spreading* (not just speaking) misinformation during a natural disaster or public health crisis? Such considerations might involve new conceptions of libel and other laws without changing basic Section 230 protections. Transparency and due process in adjudication of penalties would also be required. For example, users deserve some rights to seek recourse if they are falsely accused of *spreading* misinformation. Enumerating user rights and responsibilities in this process would be a useful and appropriate use of regulator effort, enabling the public to shape, through regulation, the contours of platform response.

Objection 2—Chilling Effect on Users

Another objection is that, by requiring users to attest to their content for it to spread, attestation will have a "chilling effect" on the free, open discourse that has important social value. However, first, and most importantly, attestation does not restrict *speech* in any way. Anyone can continue to say/post whatever they choose without liability. What attestation restricts is the *audience* for their speech. In attestation, the audience is not "the public," but only those users who have declared a relationship with the speaker, such as by "following/friending" them. It is only the intention to go beyond this audience that triggers responsibility. There is no need to take on additional liability for an individual user to, for example, share speculative celebrity gossip with their friends (Chadwick, Vaccari, & Loughlin, 2018). Alternatively, users may elect to attest to content they think is worthy of public attention, but only do so when they are confident that what they are posting is accurate.

Empirically, liability because of attesting for most users will be limited as most users do not share fake news (Guess, Nagler, & Tucker, 2019), likely because many users' friends and close ties still hold them accountable for misinformation (Bode & Vraga, 2018). Thus, for these users, the loss of the theoretical ability to spread content they would not stand behind to large audiences they have no prior relationship with is a minimal concern.

Attestation also does not restrict users who already broadcast accurate information and are willing to attest to it. This case applies to "accountable spreaders" in Table 1. There are many kinds of professional users for whom institutional norms and reputational costs encourage them to spread only accurate information. For example, many journalists' messages are already published directly by media platforms that bear legal responsibility for their content. Thus, the journalist's work is *already attested to* as the costs

of attestation are largely already borne. Many journalists also use social media to convey information that is more speculative and not yet publishable by journalistic standards. Importantly, since attestation is a decision at the post level, not the author level, journalists need not attest to every post they make. Their attestation decision then sends an important signal that distinguishes these posts.

Another category of users that may appear to be adversely affected by attestation are social activists. Social media have been credited with facilitating the rise of many social movements that gave suppressed voices the ability to broadcast to the wider public (Jackson & Foucault Welles, 2016; Papacharissi & de Fatima Oliveira, 2012). It may appear that attestation relegates these voices once again to the background. However, this appearance is misleading.

First, the potential for these movements cannot be evaluated by considering them alone; it must also account for their competition, both for attention and from explicit countermovements (Freelon, McIlwain, & Clark, 2018; Zhang et al., 2019) and "astro-turfers" (Keller, Schoch, Stier, & Yang, 2020; Zerback, Töpfl, & Knöpfle, 2021). While attestation may, in some circumstances, slow the spread of accurate messages that identify injustice on a topic, it should reduce the spread of "astro-turfing" and other industrial spread seekers to a greater degree. For example, Tufekci (2017) documents how intensive, personal interactions and organization were critical to the civil rights movement. Such efforts would not require anyone to attest. By contrast, movements that lack connections to a meaningful underlying social network of real people would be stymied, as, absent attestation, they would only be heard within their own (artificial) echo chamber.

It also must be noted that, in recent years, social movements have increasingly emerged based on false information, such as the #stopthesteal movement that culminated in the events of January 6. This adaptation follows the basic logic of unaccountable spread seeking. If social media is an effective way to organize political voices, and this organization is completely uninhibited by any promise to be truthful, then bad-faith actors will seek to create movements based on false information. In other words, although attestation does, in general, put an additional burden on all social movements, it puts a much greater burden on artificial movements and those organized on false premises, leaving more attention for the others.

Objection 3—Non-Deterrence to Bad-Faith Users With Large Followings

A third objection is that attestation would not address the problem of the spread of harmful messages by actors with large followings, such as politicians, media personalities, or other "bad-faith" celebrities. It is true that within this framework such individuals can, in principle, spread harmful messages, including misinformation, to their audiences without attesting and thus evade liability. This may seem unjust in that attestation appears to constrain "the little guy" more than the most consequential offenders.

However, adoption of attestation, although not immediately restricting bad-faith celebrities, would nonetheless have some salutary effects on their behavior. First, they would retain their safety only by refraining from attesting, and this refusal to attest is still broadcast to anyone who accesses that post, nonetheless. It thus creates an explicit declaration that they do not stand behind the content they have posed. This explicit declaration would serve as a clear target for criticism for journalists or others who wished to hold them accountable. For example, rather than having to first do research to fact-check the statement,

they could immediately publicize reasons to doubt it by factually reporting that the speaker themselves "refused to stand behind it." Second, the framework's incentives would reduce these individuals' abilities to cheaply access false claims. Currently, false claims made by anyone can quickly diffuse to anyone else. This creates a "crowdsourced" pipeline of misinformation that bad-faith celebrities can simply reshare with one click (Bovet & Makse, 2019). Attestation should substantially reduce this supply, since now the celebrity would have to, him- or herself, access this supply through direct relationships, constricting (though not eliminating) it. Third, by reducing the quantity of misinformation produced in general, norms of platform behavior could shift. In particular, under attestation, hashtags and trending lists would be (mostly) clean, such that the false claims shared by bad-faith celebrities may appear more out of place rather than just part of the nature of the platform.

All of this said, the problem of reducing the harms created by the speech of individuals who are well known and have large, dedicated audiences is difficult to address with any particular policy. That attestation does not solve this highly intractable problem should not be held against it.

Discussion and Conclusion

A fundamental flaw in the design of social media is the failure to anticipate the opportunities it offers unaccountable spread seekers. In a media regime in which accounts are easy to create, posts can be cheaply produced at scale, and there is no responsible party attached to any message, those seeking to wield influence without compunction have a clear strategy: Flood the system with messages that serve their interest, regardless of the harms they cause.

Analysis of this problem has been further inhibited by a conceptual blurring between the "right to speak," to express oneself freely, and the ability to "spread," to have one's speech heard by the public at large. Attempts to regulate online discourse thus appear to run afoul of First Amendment concerns (in spirit if not by law), and so the public consequences of harmful speech, while recognized in practice, are not given due weight in principle. Prior media regimes recognized that although the right to express oneself freely is important, the ability to abuse this right for the sake of spreading one's speech to large audiences should be a privilege that must be earned and can be revoked. New technologies removed this relationship in a technical sense, so it falls to the regulatory regime to restore it without undoing the benefits, particularly in open access to more voices these technologies have provided.

We have suggested attestation as one way to restore this balance between freedom to speak and accountability for speech within the affordances and flexibility of online communication. Attestation draws a bright line between speech and spread and allows speakers autonomy in choosing the intent for their speech and thus the responsibilities they undertake for it. In so doing, it also acknowledges that, though platforms have enormous data and resources, there is much they do not know about what users are saying in the millions of messages posted each day, and some of this information the user *does* know. Thus, although a sense of justice may point toward holding platforms responsible, as they are the ones who reap the benefits of user actions, this moral assessment overlooks the practical fact that users must be encouraged to participate responsibly, too.

We have proposed one way of achieving this through the attestation principle and have illustrated how this simple principle would facilitate reducing misinformation and other harmful messages on social media while having minimal impact on most good-faith users. We do not see this proposal as an end to the discussion but as an initiation of further conversation that takes a realistic and systemic view of the problem.

References

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining—WSDM '11* (pp. 65–74). Hong Kong, China: ACM Press. doi:10.1145/1935826.1935845

Beck, T., Bühren, C., Frank, B., & Khachatryan, E. (2020). Can honesty oaths, peer interaction, or monitoring mitigate lying? *Journal of Business Ethics, 163*(3), 467–484. doi:10.1007/s10551-018-4030-z

Bernays, E. (2020). *Propaganda* (C. Clarke, Ed.). Cupertino, CA: AIOS Publishing.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives, 12*(3), 151–170. doi:10.1257/jep.12.3.151

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication, 33*(9), 1131–1140. doi:10.1080/10410236.2017.1331312

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications, 10*(7), 1–14. doi:10.1038/s41467-018-07761-2

Burt, R. S. (2005). *Brokerage and closure: An introduction to social capital*. New York, NY: Oxford University Press.

Busby, M. (2018, April 17). You can buy anything on the black market—including Twitter handles. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/apr/17/selling-twitter-handles-big-business-identity

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science, 3*(1), 14–25. doi:10.1002/bs.3830030103

Cecka, D. M. (2014). Abolish anonymous reporting to child abuse hotlines. *Catholic University Law Review, 64*(1), 51–98.

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media & Society, 20*(11), 4255–4274. doi:10.1177/1461444818769689

Chen, A. (2020, September 26). Limiting message forwarding on WhatsApp helped slow disinformation. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/2019/09/26/434/whatsapp-disinformation-message-forwarding-politics-technology-brazil-india-election/

Cheng, J., Adamic, L. A., Kleinberg, J. M., & Leskovec, J. (2016). Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web* (pp. 671–681). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. doi:10.1145/2872427.2882993

Coviello, L., Sohn, Y., Kramer, A. D. I., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). Detecting emotional contagion in massive social networks. *PLoS One, 9*(3), 1–6. doi:10.1371/journal.pone.0090315

De Bruin, B. (2016). Pledging integrity: Oaths as forms of business ethics management. *Journal of Business Ethics, 136*(1), 23–42. doi:10.1007/s10551-014-2504-1

Facebook MAU worldwide 2022. (2022). Statista. Retrieved from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Feldman, Y. (2018). *The law of good people: Challenging states' ability to regulate human behavior*. Cambridge, UK: Cambridge University Press.

Freelon, D., McIlwain, C., & Clark, M. (2018). Quantifying the power and consequences of social media protest. *New Media & Society, 20*(3), 990–1011. doi:10.1177/1461444816676646

Fraud & abuse laws. (2021, September 1). Office of Inspector General | Government Oversight | U.S. Department of Health and Human Services. Retrieved from https://oig.hhs.gov/compliance/physician-education/fraud-abuse-laws/

Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text*, (25/26), 56–80. https://doi.org/10.2307/466240

Frey, S., Albino, D. K., & Williams, P. L. (2018). Synergistic information processing encrypts strategic reasoning in poker. *Cognitive Science, 42*(5), 1457–1476. doi:10.1111/cogs.12632

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (Vol. 167, pp. 167–194). Cambridge, MA: MIT Press.

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2016). The structural virality of online diffusion. *Management Science, 62*(1), 180–196. doi:10.1287/mnsc.2015.2158

Gordon, M. D. (1980). The perjury statute of 1563: A case history of confusion. *Proceedings of the American Philosophical Society, 124*(6), 438–454.

Green, S. P. (2001). Lying, misleading, and falsely denying: How moral concepts inform the law of perjury, fraud, and false statements. *Hastings Law Journal, 53*(1), 157–212.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*(6425), 374–378. doi:10.1126/science.aau2706

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(1), eaau4586. doi:10.1126/sciadv.aau4586

Guinaudeau, B., Munger, K., & Votta, F. (2022). Fifteen seconds of fame: TikTok and the supply side of social video. *Computational Communication Research, 4*(2), 463–485. doi:10.5117/CCR2022.2.004.GUIN

Hannan, M. T., Polos, L., & Carroll, G. R. (2007). *Logics of organization theory: Audiences, codes, and ecologies*. Princeton, NJ: Princeton University Press.

Harvard Law Review. (2018). Section 230 as First Amendment Rule. *Harvard Law Review, 131*(7), 2027–2048.

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., & Yu, C. (2015). The quest to automate fact-checking. In *Proceedings of the 2015 Computation+ Journalism Symposium* (pp. 1–5). New York, NY. Retrieved from http://cj2015.brown.columbia.edu/papers.html

Instagram. (n.d.). *Create a new Instagram account*. Instagram. Retrieved October 25, 2024, from https://help.instagram.com/155940534568753

Jackson, S. J., & Foucault Welles, B. (2016). #Ferguson is everywhere: Initiators in emerging counterpublic networks. *Information, Communication & Society, 19*(3), 397–418. doi:10.1080/1369118X.2015.1106571

Kain, E. (2018, January 18). YouTube is demonetizing small channels, and that's a good thing. *Forbes*. Retrieved from https://www.forbes.com/sites/erikkain/2018/01/18/youtube-is-demonetizing-small-channels-and-why-thats-a-good-thing/

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication, 37*(2), 256–280. doi:10.1080/10584609.2019.1661888

Keller, T. R., & Kleinen-von Königslöw, K. (2018). Followers, spread the message! Predicting the success of Swiss politicians on Facebook and Twitter. *Social Media + Society, 4*(1), 1–11. doi:10.1177/2056305118765733

Kleinnijenhuis, J., van Hoof, A. M. J., & van Atteveldt, W. (2019). The combined Effects of mass media and social media on political perceptions and preferences. *Journal of Communication, 69*(6), 650–673. doi:10.1093/joc/jqz038

Lawson, M. A., & Kakkar, H. (2022). Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. *Journal of Experimental Psychology: General, 151*(5), 1154–1177. doi:10.1037/xge0001120

Lazer, D. (2015). The rise of the social algorithm: Does content curation by Facebook introduce ideological bias? *Science, 348*(6239), 1090–1091. doi:10.1126/science.aab1422

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. doi:10.1126/science.aao2998

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '09* (pp. 497–506). Paris, France: ACM Press. doi:10.1145/1557019.1557077

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131. doi:10.1177/1529100612451018

MacCormick, N. (1983). What is wrong with deceit. *Sydney Law Review, 10*(1), 5–19.

Marwick, A. E., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133. doi:10.1177/1461444810365313

McTeague, M. (1988). Neither rain, nor sleet . . . nor the United States Congress . . . will prevent the U.S. Postal Service from delivering Hustler Magazine. *Loyola Entertainment Law Journal, 8*(1), 159–168.

Meta. (n.d). *Penalties for sharing fact-checked content*. Meta. Retrieved October 25, 2024, from https://transparency.meta.com/enforcement/taking-action/penalties-for-sharing-fact-checked-content/

O'Sullivan, P. B., & Carr, C. T. (2017). Masspersonal communication: A model bridging the mass-interpersonal divide. *New Media & Society, 20*(3), 1161–1180. doi:10.1177/1461444816686104

Papacharissi, Z., & de Fatima Oliveira, M. (2012). Affective news and networked publics: The rhythms of news storytelling on #Egypt. *Journal of Communication, 62*(2), 266–282. doi:10.1111/j.1460-2466.2012.01630.x

Patel, F., & Hecht-Felella, L. (2021, February 22). *Facebook's content moderation rules are a mess*. Brennan Center for Justice. Retrieved from https://www.brennancenter.org/our-work/analysis-opinion/facebooks-content-moderation-rules-are-mess

Paul, C., & Matthews, M. (2016). The Russian "firehose of falsehood" propaganda model: Why it might work and options to counter t. *RAND Corporation, 2*(7), 1–10. doi:10.7249/PE198

Peer, E., & Feldman, Y. (2021). Honesty pledges for the behaviorally-based regulation of dishonesty. *Journal of European Public Policy, 28*(5), 761–781. doi:10.1080/13501763.2021.1912149

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(7855), 590–595. doi:10.1038/s41586-021-03344-2

Rainie, H., & Wellman, B. (2012). *Networked: The new social operating system*. Cambridge, MA: MIT Press.

Securities fraud laws. (2019, May 28). Justia. Retrieved from https://www.justia.com/criminal/offenses/white-collar-crimes/securities-fraud/

Shteynberg, G. (2015). Shared attention. *Perspectives on Psychological Science, 10*(5), 579–590. doi:10.1177/1745691615589104

Smith v. California, 361 U.S. 147 (1959). (1959). Justia Law. Retrieved from https://supreme.justia.com/cases/federal/us/361/147/

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon bombing. In *iConference 2014 Proceedings* (pp. 1–9). Urbana-Champaign, IL: iSchools. doi:10.9776/14308

Subramanian, S. (2017, February 15). Meet the Macedonian teens who mastered fake news and corrupted the U.S. election. *Wired*. Retrieved from https://www.wired.com/2017/02/veles-macedonia-fake-news/

Sutcliffe, A., Dunbar, R., Binder, J., & Arrow, H. (2012). Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology, 103*(2), 149–168. doi:10.1111/j.2044-8295.2011.02061.x

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication, 33*(3), 460–480. doi:10.1080/10584609.2015.1102187

TikTok. (2022, September 28). *An update on our work to counter misinformation*. TikTok. Retrieved from https://newsroom.tiktok.com/en-us/an-update-on-our-work-to-counter-misinformation

TikTok. (n.d.). *Sign up for TikTok*. TikTok. Retrieved October 25, 2024, from https://www.tiktok.com/signup

Tobin, A., Varner, M., & Angwin, J. (2017, December 28). *Facebook's uneven enforcement of hate speech rules allows vile posts to stay up*. ProPublica. Retrieved from https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes?token=2IGQdWFIQ6nSzBbnQ4-16RVFd3ayz6DV

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. New Haven, CT: Yale University Press.

Twitter Usage Statistics—Internet Live Stats. (2022, June 15). Retrieved from https://www.internetlivestats.com/twitter-statistics/

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*, 280–289. Montreal, Canada: AAAI Publications. https://doi.org/10.1609/icwsm.v11i1.14871

Varol, O., & Uluturk, I. (2018). Deception strategies and threats for online discussions. *First Monday, 23*(5–7). doi:10.5210/fm.v22i5.7883

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior, 32*(5), 375–389. doi:10.1007/s10979-007-9116-6

X. (n.d.). *How to sign up for an X account*. X. Retrieved from https://help.x.com/en/using-x/create-x-account

X. (2023, April). *Synthetic and manipulated media policy*. X. Retrieved October 25, 2024, from https://help.x.com/en/rules-and-policies/manipulated-media

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies, 1*(1), 48–61. doi:10.1002/hbe2.115

Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Who let the trolls out?: Towards understanding state-sponsored trolls. In *Proceedings of the 10th ACM Conference on Web Science—WebSci '19* (pp. 353–362). Boston, MA: ACM Press. doi:10.1145/3292522.3326016

Zerback, T., Töpfl, F., & Knöpfle, M. (2021). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media & Society, 23*(5), 1080–1098. doi:10.1177/1461444820908530

Zhang, Y., Shah, D., Foley, J., Abhishek, A., Lukito, J., Suk, J., . . . Garlough, C. (2019). Whose lives matter? Mass shootings and social media discourses of sympathy and policy, 2012–2014. *Journal of Computer-Mediated Communication, 24*(4), 182–202. doi:10.1093/jcmc/zmz009

Zittrain, J. (2014, June 1). Facebook could decide an election without anyone ever finding out. *The New Republic*. Retrieved from https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering