Article

SM+S social media + society

Social Media + Society
January-March 2022: 1–13
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20563051211069048
journals.sagepub.com/home/sms
SAGE

Fact-Checking the Crisis: COVID-19, Infodemics, and the Platformization of Truth

Kelley Cotter[1], Julia R. DeCook[2], and Shaheen Kanthawala[3]

Abstract

During the onset of the COVID-19 pandemic, various officials flagged the critical threat of false information. In this study, we explore how three major social media platforms (Facebook, Twitter, and YouTube) responded to this "infodemic" during early stages of the pandemic via emergent fact-checking policies and practices, and consider what this means for ensuring a well-informed public. We accomplish this through a thematic analysis of documents published by the three platforms that address fact-checking, particularly those that focus on COVID-19. In addition to examining what the platforms said they did, we also examined what the platforms actually did in practice via a retrospective case study drawing on secondary data about the viral conspiracy video, *Plandemic*. We demonstrate that the platforms focused their energies primarily on the visibility of COVID-19 mis/disinformation on their sites via (often vaguely described) policies and practices rife with subjectivity. Moreover, the platforms communicated the expectation that users should ultimately be the ones to hash out what they believe is true. We argue that this approach does not necessarily serve the goal of ensuring a well-informed public, as has been the goal of fact-checking historically, and does little to address the underlying conditions and structures that permit the circulation and amplification of false information online.

Keywords

fact-checking, misinformation, disinformation, platforms, COVID-19

In February 2020, the Director-General of the World Health Organization (WHO) cautioned that the spread of "fake news" had the potential to exacerbate the spread of COVID-19, declaring "we're not just fighting an epidemic; we're fighting an infodemic" (Ghebreyesus, 2020). One month later, the Associate Director of the International Fact-checking Network (IFCN) described COVID-19 as "the biggest challenge fact-checkers have ever faced" (Suárez, 2020). By this point, Facebook and YouTube had begun to develop partnerships with fact-checking organizations, and fact-checking gradually became a go-to response for managing information pollution (misinformation, disinformation, malinformation) (Wardle & Derakhshan, 2017). Because platforms are key vectors for misinformation, particularly pandemic-related misinformation (Bridgman et al., 2020), ensuring that the global public are well-informed requires assurance that platform fact-checking functions effectively and in the public's best interest. In a very real sense, the global population's safety and wellness hinges on the strength and quality of platforms' regimes of fact-checking.

More broadly, platforms' role in and response to infodemics symptomize the rise of the "platform society"—how platforms have "penetrated the heart of societies—affecting institutions, economic transactions, and social and cultural practices" (van Dijck et al., 2018, p. 2). In this state of affairs, platforms' private interests and values do not always perfectly match those of the diversity of publics that global platforms like Facebook, Twitter, and YouTube serve. Moreover, there is an ongoing debate about what rights and responsibilities platforms have in relation to user generated content and activity (e.g., Gillespie, 2018b; Gorwa, 2019; van Dijck et al., 2018). Early reports suggest that platform fact-checking programs invoke many longstanding questions and concerns around "governance by platforms," or the policies, design choices, and business models that structure participation on

[1]Pennsylvania State University, USA
[2]Loyola University Chicago, USA
[3]University of Alabama, USA

Corresponding Author:
Kelley Cotter, College of Information Sciences and Technology, Pennsylvania State University, E331 Westgate Building, University Park, PA 16802, USA.
Email: kcotter@psu.edu

these sites (Gillespie, 2018b), particularly those related to content moderation.

The purpose of this study is to explore how platforms responded to the infodemic during early stages of the pandemic via emergent fact-checking policies and practices, particularly in the United States. We focus on how three major platforms (Facebook, Twitter, and YouTube) turned to and integrated their governance structures with the existing infrastructure of fact-checking, which has traditionally been the purview of journalists. In this way, we connect literatures on mis/disinformation and fact-checking with the growing body of work on platform governance to consider how platforms are reconfiguring fact-checking and with what effect. To accomplish this, we conducted a thematic analysis of official documents published by Facebook, Twitter, and YouTube that address fact-checking and focus on COVID-19. In addition to examining what the platforms said they did, we also examined what the platforms actually did in practice via a retrospective case study drawing on secondary data about the viral video *Plandemic*, which propagated various false claims and conspiracy theories about COVID-19.

Acknowledging that platform policies change rapidly, we note that this research focuses on platform policies and (in) actions that occurred early on during the COVID-19 pandemic (from March 2020 until November 2020, prior to vaccine availability). As such, this research documents an important historical moment in platforms' nascent practices and policies around fact-checking and COVID-19-related mis/disinformation.

Overall, our findings reveal a familiar emphasis on vague and subjective policies and practices for limiting the *visibility* of mis/disinformation. As in past approaches to content moderation, the platforms communicate the expectation that users should ultimately be the ones to hash out what they believe is true. We argue that this deviates from the goal of fact-checking, traditionally, which is to ensure a well-informed public.

Infodemics and Information Pollution

The struggle to combat the spread of false information has become an era-defining issue. The loss of trust in institutions (media, government, education, and in our context, health organizations) in recent years has contributed to the pervasiveness and high levels of susceptibility to false information (Humprecht et al., 2020; Swire-Thompson & Lazer, 2020). Moreover, polarization, populism, fragmentation, shifts toward the online engagement-based advertising ecosystem, and the decline of local journalism have collectively established conditions conducive to "information pollution" (Humprecht et al., 2020; Wardle & Derakhshan, 2017).

Information pollution serves as an umbrella term for different kinds of problematic information. This article focuses on two main forms of information pollution: misinformation and disinformation. Misinformation refers to when people share unsubstantiated claims, rumors, and conspiracy theories that they do not recognize as untrue or inaccurate. Disinformation refers to when people *intentionally* share false information to propagate a certain point of view by deceiving audiences.

In the health context, information pollution is especially dangerous. Studies have consistently demonstrated how mis/disinformation have contributed to the spread of diseases (Gyenes & Mina, 2018), even ones previously thought eradicated (Swire-Thompson & Lazer, 2020). Indeed, early evidence suggests that the preexisting anti-vaccine propaganda has contributed to global reluctance to receive a COVID-19 vaccine (Smith et al., 2020). In the domain of health and science, mis/disinformation is defined in relation to contemporaneous expert consensus (Tan et al., 2015) and "best available evidence" (Garrett et al., 2016 p. 333). That is, a claim can be considered false if most experts agree and the preponderance of scientific evidence suggests that it is false (Vraga & Bode, 2020).

In online settings, false information often spreads *farther* and *faster* than facts (Ball & Maxmen, 2020; Lewandowsky et al., 2017; Swire-Thompson & Lazer, 2020). As such, information pollution has increased in scale and severity in recent years, as false information can quickly spread from platform to platform, evading moderation (Ferrara et al., 2020; Gyenes & Mina, 2018) and making it difficult to control. The mechanisms by which false information spreads and who accepts it are still under constant discovery. Existing work suggests that platforms' technical infrastructures—particularly, algorithms that sort, filter, and recommend content—play a significant role in connecting people with digital spaces and conversations generative of information pollution (Gillespie, 2020; Kaiser et al., 2021; Villasenor, 2020). This state of affairs has placed new demands on platforms to engage in fact-checking.

The Evolution of Fact-Checking

Fact checking has evolved over time, and currently plays a key role in the so-called "post-truth" media landscape. Historically, fact-checking has been associated with journalism, reflecting its professionalization in the 20th century as a "fact-centered discipline" (Graves & Amazeen, 2019, p. 3). Yet, over time, fact-checking has developed into a broader infrastructure composed of an eclectic mix of people and organizations, practices and routines, principles, and tools interconnected in support of the goal of "helping people become better informed and promoting fact-based public discourse" (Graves & Amazeen, 2019, p. 1).

In the early 21st century, fact-checking significantly expanded and began to revolve around ensuring institutional accountability (Graves & Amazeen, 2019). This shift likely grew in response to a series of claims in the political sphere that proved false, with the Bush Administration's claims about weapons of mass destruction as justification for the

Iraq War as a watershed moment (Marietta et al., 2015). Simultaneously, the Internet democratized the ability to produce news-like content. This allowed independent, and sometimes amateur, fact-checking sites (e.g., Snopes.com) to establish themselves to help dispel conspiracy theories and rumors while also striving to serve as watchdogs for politicians, journalists, and other public figures (Graves & Amazeen, 2019). Soon, some major media organizations developed fact-checking arms, in response to demand for greater oversight of political campaigns and government (Graves & Amazeen, 2019).

While media organizations dominate fact-checking, comprising more than half of all fact-checking organizations, their role has weakened over time with increasingly more nonprofits, think tanks, nongovernmental organizations, and academic institutions joining the ranks (Bell, 2019; Stencel & Luther, 2021). The extraordinary levels of misinformation and disinformation in the 2016 U.S. election, further engendered demand for fact-checking with a resultant 200% increase in fact-checking organizations (Fischer, 2020). According to the Duke Reporters' Lab census, as of June 2021, 341 fact-checking organizations are active worldwide (Stencel & Luther, 2021). Demonstrating fact-checking's evolution into its own subindustry, in 2015, the IFCN (n.d.) was formed, and a year later the organization introduced a Code of Principles, which prioritized transparency, nonpartisanship, and fairness.

Fact-checking is now considered an essential tool for combating false information, particularly online. While past work found that fact-checking corrections can create a backfire effect, and reinforce people's original, inaccurate beliefs (Nyhan & Reifler, 2010; Nyhan et al., 2013), more recent research has shown fact-checking to be successful in improving the accuracy of beliefs, supporting the ability to correctly evaluate claims, and reducing intentions to share untrue headlines on social media (Amazeen et al., 2015; Nyhan et al., 2020; Porter & Wood, 2020; Yaqub et al., 2020). Together, these findings suggest that fact-checking can effectively fulfill its core aim of ensuring a well-informed public, which makes it an important tool for online platforms as they have become key venues for keeping abreast of news and current events.

Fact-Checking in the Platform Era

After the 2016 U.S. presidential election, concerns about rampant mis/disinformation online escalated, and major U.S.-based platforms began to implement their own fact-checking programs in response (Ananny, 2018). As will be discussed, Facebook and YouTube have built relationships with third-party fact-checking organizations and Twitter has handled fact-checking internally. These fact-checking programs represent an extension of platforms' "apparatus for content moderation" (Gillespie, 2020, p. 329), in which they have invested considerable resources over the past decade. Content moderation is a core mechanism by which platforms govern activity on their sites, shaping "what is seen, what is valued, what is said" (Grimmelmann, 2015, p. 42).

By looking to existing work on content moderation, we can see how the platform fact-checking program may differ from pre-existing fact-checking efforts. First, content moderation follows a particular politico-economic logic that may conflict with arbitrating truths. Content is platforms' core commodity (Roberts, 2018), and public outcry and legal pressure have made it good business for them to address various kinds of "problematic" content on their sites. Content moderation policies and practices should be understood as "compromises—between users with different values and expectations, as well as between the demands of users and the demands of profit" (Gillespie, 2018a, p. 12). Yet, in such compromises, the values, expectations, and demands of certain users and stakeholders matter more than others. Platforms commonly rely on "tiered governance," exempting high-profile accounts from normal enforcement of policies or subjecting them to a more lenient set of rules (Caplan & Gillespie, 2020; Horwitz, 2021). For example, under pressure from conservative politicians and advocacy groups, Facebook has urged fact-checking partners to alter "False" determinations, particularly for prolific advertisers (Pasternack, 2020).

This politico-economic logic of content moderation positions it at odds with fact-checking's goal of achieving institutional accountability. Historically, platforms like Facebook, Twitter, and YouTube have been loath to "choose sides," which in some respects institutional accountability presupposes. Instead, they aim to give the appearance of noninterventionist mediators (Gillespie, 2018a), "empowering all by choosing none" (Gillespie, 2010, p. 357). Platforms lean on a cyberlibertarian approach that places the onus on individual users, deemed rational actors, to work out the truth or the "best" ideas through competition in the "marketplace of ideas" (Maddox & Malson, 2020).

Second, content moderation on platforms depends upon both people and artificial intelligence (AI), but, given the global scale of content, the latter is used to detect the majority of cases and to enforce rules (Gillespie, 2018a). Algorithmic moderation is often perceived and positioned as more objective than human judgments (Gillespie, 2018a; Roberts, 2018). Yet, algorithmic moderation is rife with errors, particularly in situations in which drawing the line between acceptable and unacceptable depends upon nuanced understanding and interpretation of sociocultural context (Gillespie, 2018a). For example, fact-checking and news organizations typically fact-check a much wider range of content than platforms, including content that can be difficult for AI to evaluate, like satire, opinion pieces, and political advertisements (Stewart, 2021). Misinformation and disinformation pose a significant problem for automated detection, as they often require making complex judgments about facts that carry partial truths, involve moral concepts, or lack consensus (Stewart, 2021). Moreover, mis/disinformation differs from other kinds of problematic content for which content moderation systems were built. Unlike porn, gore, or the illegal sale of regulated goods, mis/disinformation is "designed to emulate exactly what the platform wants to distribute most" (Gillespie, 2020, p. 334). Although AI is gradually being integrated into the fact-checking infrastructure, some suggest human oversight remains key (Adair & Stencel, 2020; Graves, 2018).

Research Questions

As platforms endeavor to develop fact-checking programs, they must adapt the preexisting fact-checking infrastructure to their goals, values, and systems. In the remainder of the article, we explore this, as it has unfolded in the context of health mis/disinformation. We ask, in the early stages of the pandemic, how have platforms responded to COVID-19 and the larger infodemic crisis via policies and practices around fact-checking, and how do they position themselves within the broader infrastructure of fact-checking? By exploring these questions, we aim to contribute to discussion of how platforms are (re)shaping our information infrastructure, which has significant implications for global resilience to information pollution (Humprecht et al., 2020), particularly amid the high stakes context of the COVID-19 pandemic. How platforms enter into and alter the nature of fact-checking matters for how we arbitrate truth in the global public sphere—who is responsible for this work and how we should operationalize it.

Method

To answer the research question above, we conducted a thematic analysis of official English-language documents issued by Facebook, Twitter, and YouTube that address fact-checking, focusing on documents that mention COVID-19. To identify documents, we searched each platform's relevant webpages (e.g., about.fb.com, blog.twitter.com, support.google.com) for the keywords "fact checking" or "fact check" and related keywords like "misinformation" and "fake news." Our search yielded blog posts, press releases, policy documents, and help pages, all representing a subset of a larger collection of documents that concern misinformation and fact-checking on the platforms in general. We cut off data collection in November 2020 (providing insights into the platforms' actions during the early days of the pandemic), resulting in a total of 312 documents. We then reviewed all the documents from January 2020 on, searching for keywords related to COVID-19 (COVID-19, coronavirus, pandemic). Our final dataset describing policies and practices related to the early days of the COVID-19 pandemic totaled 60 documents (see supplementary materials for a list of documents). Of note, all three authors are based in the United States, which meant that the platforms defaulted to showing policies in English and specifically for U.S.-based users.

We qualitatively coded these 60 documents via an iterative process, focusing on what the platforms said about how they were responding to the COVID-19 infodemic and how they rhetorically positioned themselves in this. First, each of our three-member research team independently coded 5 different documents each (i.e., 15 documents). We discussed codes, occasionally merging and synthesizing higher-level codes. After, we each coded six additional documents and discussed and adjusted codes again (18 documents). Finally, we divided the remaining 27 documents among us (9 per researcher) and coded them independently. At this point, we met to further develop more abstract themes. The final codebook included codes highlighting fact-checking actions taken by the platforms (including codes about third-party fact-checking partnerships, removal of content based on fact-checking, and encouraging content creators to do their own fact-checking) and their enforcement strategies used to identify and prevent the circulation of problematic information (such as limiting visibility through warning labels, demoting, or removal of content).

We also conducted a retrospective case study of the viral video *Plandemic*, which contained a series of false claims about COVID-19 and public health officials. Drawing on secondary sources, mainly media reports, we sought to understand (1) what actions platforms took in relation to the *Plandemic* videos, (2) when they took action, and (3) why they said they took action. We used this case study as a tool for reflecting on how platforms actually applied their policies and practices in practice.

Findings

In the three platforms' descriptions of their policies and practices, we can see how they have integrated themselves within the infrastructure of fact-checking, treating it as special area of content moderation. At the highest level, the platforms described a twofold response to COVID-19 mis/disinformation: (1) verification of information and (2) limiting the visibility of false or inaccurate information. Yet, we saw that, unlike fact-checking journalists and organizations who focus on adjudicating the veracity of claims, the platforms exhibited a principal concern over the *visibility* of potentially false and inaccurate information. Such a concern is reflective of platforms' commitment to remaining neutral and the expectation that users will resolve the truth for themselves.

Verification of Information

Fact-Checking Programs (and Who's Responsible?). Facebook's and YouTube's fact-checking programs center partnerships built with existing third-party organizations. In multiple blog posts, Facebook highlighted its work "with over 60 fact-checking organizations around the world that review content in more than 50 languages" [FB56]. These 60 (and counting) partners represent those that have been certified by IFCN. YouTube described a similar relationship with third-party fact-checkers, though specifying that any U.S. publisher could generate fact-checks, so long as they were "either a verified signatory of the International Fact Checking Network's (IFCN) Code of Principles or are an authoritative publisher" (YT21).

In contrast to Facebook and YouTube, Twitter rarely referred to "fact checking" in its communications. Indeed, only a handful of Twitter documents mention "fact check[ing]," and these refer to work by external organizations independent of (and not working directly with) the platform. However, the company described efforts that amounted to fact-checking in multiple blog posts. For example, in a blog post in May 2020, Twitter announced new measures for dealing with "misleading information," which included a section describing how the company defined and identified such tweets [TW2]. Yet, Twitter shared few other details about the internal process of fact-checking, including who performs this work. In January 2021, Twitter changed course, announcing a community-based model of fact-checking tweets called "Birdwatch."[1] While still in piloting, this tool would enroll average users in the process of fact-checking by allowing anyone to "write notes that provide informative context" (Coleman, 2021). Then, users would rate the quality and helpfulness of these notes, which would inform "future reputation models that recognize those whose contributions are consistently found helpful by a diverse set of people" (Twitter, *n*.d.). This mirrors the platform's position (and, indeed, other platforms' positions) that the "truth" should arise from discussion and debate among the user community, as the company put it "consensus from a broad and diverse set of contributors" (Twitter, n.d.).

Making Verification Sociotechnical. The three platforms bring together people and technologies—manual and automated processes—for selecting content to be fact-checked and, under some circumstances, actually performing a function akin to verification. Automated systems identify content that may include false or inaccurate claims and some of these are sent to fact-checkers and moderation staff for human review. This approach resembles the platforms' approach to moderating content in general, which revolves around similar algorithmic processes refined over several years (Gillespie, 2018a). All three platforms noted an increased reliance on automated processes as a result of reduced and remote workforces during the pandemic, although the platforms were subtle in communicating the extent of this reliance and shared very little in general about what their automated processes do and how they function.

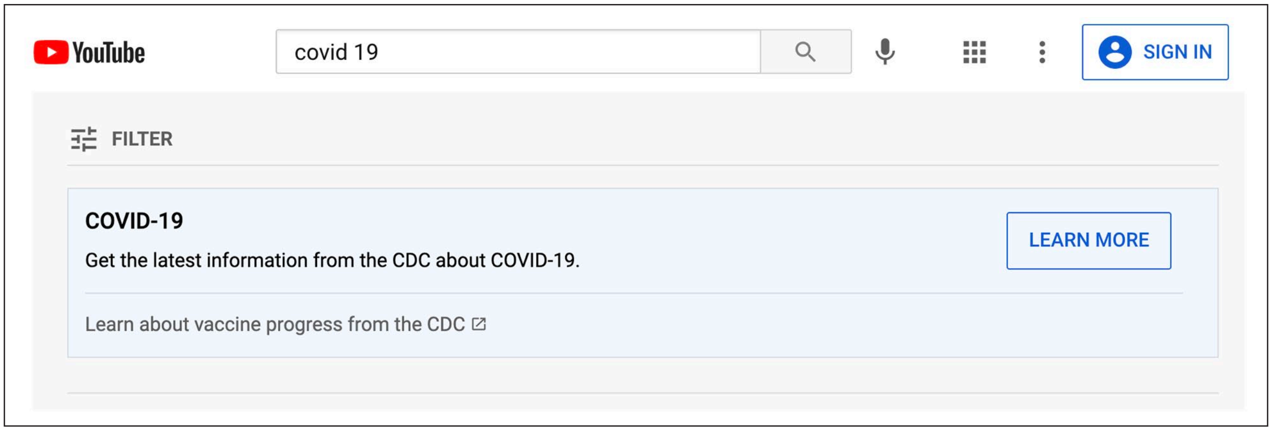
For example, Twitter explained in a blog post "Our teams are using and improving on internal systems to proactively monitor content related to COVID-19. These systems help ensure we are not amplifying Tweets with these warnings or labels and detecting the high-visibility content quickly" (TW3). Similarly, YouTube noted in a blog post: "Each quarter, millions of videos that are first flagged by our automated systems are later evaluated by our human review team and determined not to violate our policies" (YT22). Facebook was a little more forthcoming, explaining that its "machine learning model identifies potential misinformation using a variety of signals. These include comments on the post that express disbelief, and whether a post is being shared by a Page that has spread misinformation in the past."[2]

As with content moderation more generally, the platforms' communications explained that these automated processes intersect with human review. For example, Facebook noted that posts flagged as potential misinformation by its machine learning model would first be sent to paid contractors who would research claims and offer a preliminary judgment on their merit, which would then be sent to fact-checking partners to aid them in selecting stories to fact-check (FB44). Moreover, the platforms' algorithms detect potential instances of false or misleading information based in large part on user flags. YouTube has innovated in this area by introducing a "trusted flagger" program, which grants certain users special tools and privileges, including giving their flags more weight in detecting mis/disinformation.[3]

Limiting the Visibility of Mis/Disinformation

Following verification by fact-checkers, the three platforms described limiting the visibility of content determined to be false. In their communications about COVID-19 misinformation, the platforms described a multifaceted approach of removing "harmful" misinformation, demoting non-harmful misinformation, and elevating authoritative voices.

Demoting and Removing. For all three platforms, content determined to be false or inaccurate as a result of verification, is de-prioritized in feeds or removed entirely. Yet, in their communications, it was not always clear how the platforms decided which content to demote versus which to remove. Facebook stated in a blog post: "Once a post is rated false by a fact-checker, we reduce its distribution so fewer people see it" (FB51). YouTube similarly indicated, following an existing practice, they would "reduc[e] recommendations of borderline content or videos that could misinform users in harmful ways" (YT27). It was not clear from Twitter's documents whether it demoted any false or inaccurate tweets in feeds, but the company did note in a blog post from May 2020: "Our teams are using and improving on internal systems to proactively monitor content related to COVID-19. These systems help ensure we're not amplifying Tweets" that were judged to contain misleading or disputed claims (TW2). Possibly, this could mean that the platform demoted tweets in feeds.

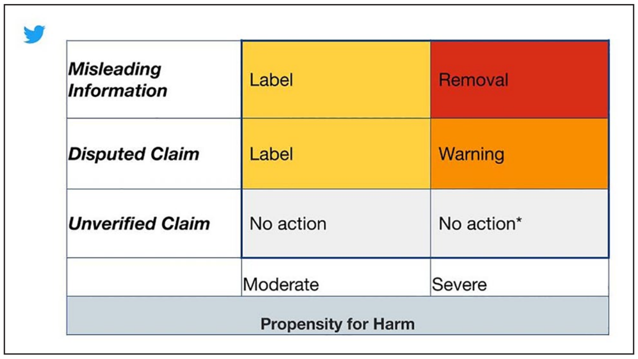


COVID-19

Get the latest information from the CDC about COVID-19.

Learn about vaccine progress from the CDC

Figure 1. YouTube COVID-19 information panel linking to the CDC website.



Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
	Propensity for Harm	

Figure 2. Twitter's table for determining how to deal with mis/disinformation (TW2).

Parallel to demoting misinformation, the platforms also emphasized their efforts to make authoritative posts more visible. For example, Twitter stated, "With a critical mass of expert organizations, official government accounts, health professionals, and epidemiologists on our service, our goal is to elevate and amplify authoritative health information as far as possible" (TW12). All three platforms also discussed dedicated spaces on their site for providing general information about COVID-19 updated in real time from health experts like the Centers for Disease Control and Prevention (CDC) and the WHO (see Figure 1), which were presented as a means of providing people with easy access to reliable sources. Notably, the platforms specified that this provision of information would not be universally available, at least initially, in all countries.

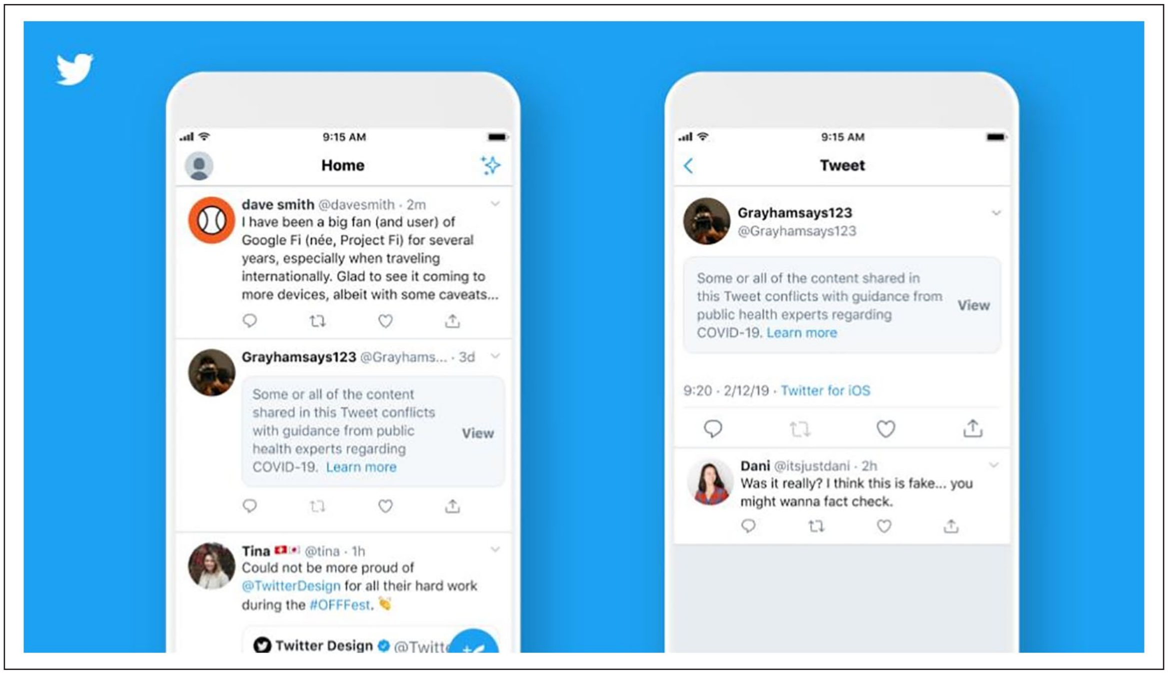
In addition to attempting to achieve the appropriate balance in the visibility of "good" and "bad" information, platforms removed some content entirely. The decision to remove content depended on the degree to which a given false claim would be considered harmful. For example, Twitter provided a table as a visual aid to explain their practices in this regard (see Figure 2). In this table, we see that the company placed "propensity for harm" as their key determining factor for applying labels, giving warnings, or removal. Namely, a tweet must contain "misleading information" and have a "severe" propensity for harm for the platform to remove it.

Similarly, Facebook stated in a blog post "We remove COVID-19 related misinformation that could contribute to imminent physical harm," and "claims that don't directly result in physical harm, like conspiracy theories about the origin of the virus" are sent to fact-checking partners to be debunked (FB48). While in one statement, YouTube referenced "harm" as a factor in reducing the distribution of "borderline" content, the company's "COVID-19 Medical Misinformation Policy" dictated that they would remove *any* content that "contradicts WHO or local health authorities' guidance on Treatment, Prevention, Diagnostic, Transmission" (YT9).

Warning Labels. Both Facebook and Twitter additionally added warning labels to posts deemed false by fact-checkers (see Figures 3 and 4). For instance, in an April 2020 blog post, Facebook explained: "Once a piece of content is rated false by fact-checkers, we reduce its distribution and show warning labels with more context" [FB51]. Recall that Twitter similarly stated that it would add warning labels to tweets that include disputed claims and have a "severe" "propensity for harm" (see Figure 2). Twitter elaborated in a blog post:

> We may label or place a warning on tweets to provide additional context in situations where the risks of harm associated with a Tweet are less severe but where people may still be confused or misled. This will make it easier to find facts and make informed decisions about what people see on Twitter. (TW3)

Although these warning labels do not make posts less visible in feeds—as in demotion and removal—in many cases, they obscure posts with an overlaid label that notifies users that a post has been "disputed." "Learn More" and "See Why"



dave smith @davesmith · 2m
I have been a big fan (and user) of Google Fi (née, Project Fi) for several years, especially when traveling internationally. Glad to see it coming to more devices, albeit with some caveats...

Grayhamsays123 @Grayhams... · 3d
Some or all of the content shared in this Tweet conflicts with guidance from public health experts regarding COVID-19. Learn more

Tina @tina · 1h
Could not be more proud of @TwitterDesign for all their hard work during the #OFFFest. 👏

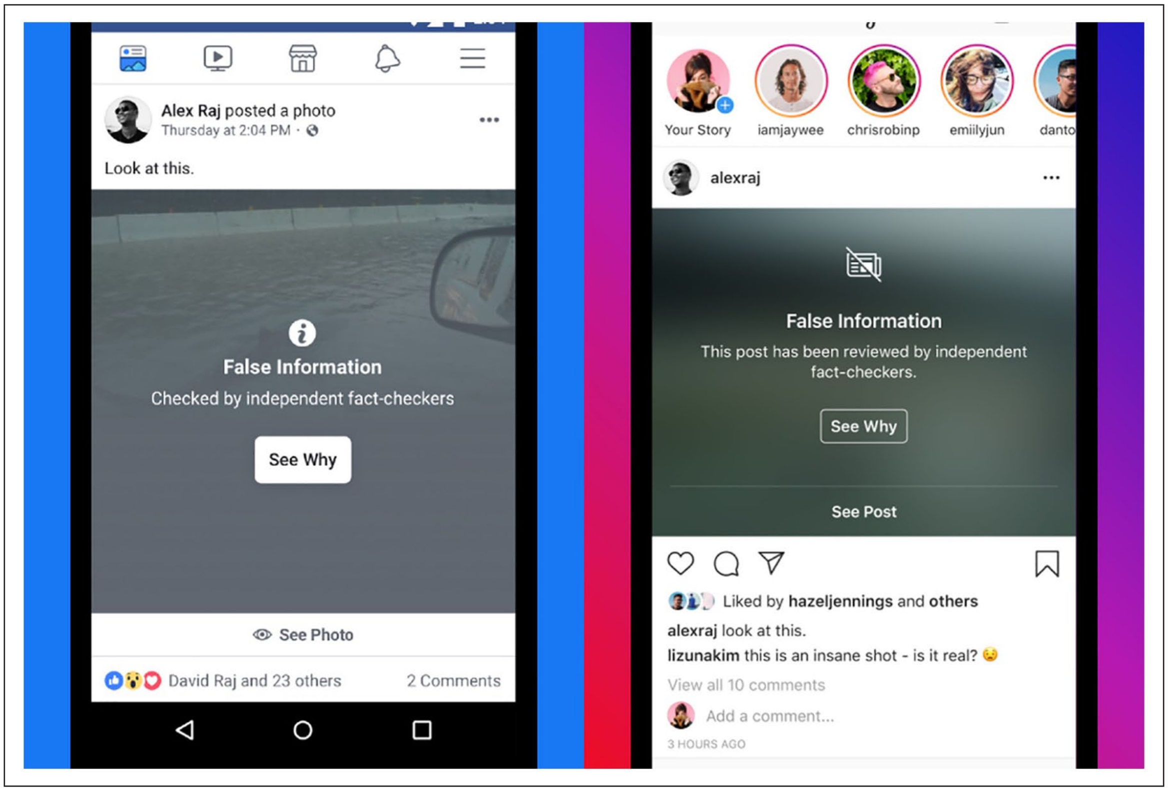
Twitter Design @Twitt

Grayhamsays123
@Grayhamsays123
Some or all of the content shared in this Tweet conflicts with guidance from public health experts regarding COVID-19. Learn more

9:20 · 2/12/19 · Twitter for iOS

Dani @itsjustdani · 2h
Was it really? I think this is fake... you might wanna fact check.

Figure 3. Twitter warning label for "disputed" COVID-19 information (TW2).



Alex Raj posted a photo
Thursday at 2:04 PM
Look at this.

False Information
Checked by independent fact-checkers

David Raj and 23 others 2 Comments

alexraj

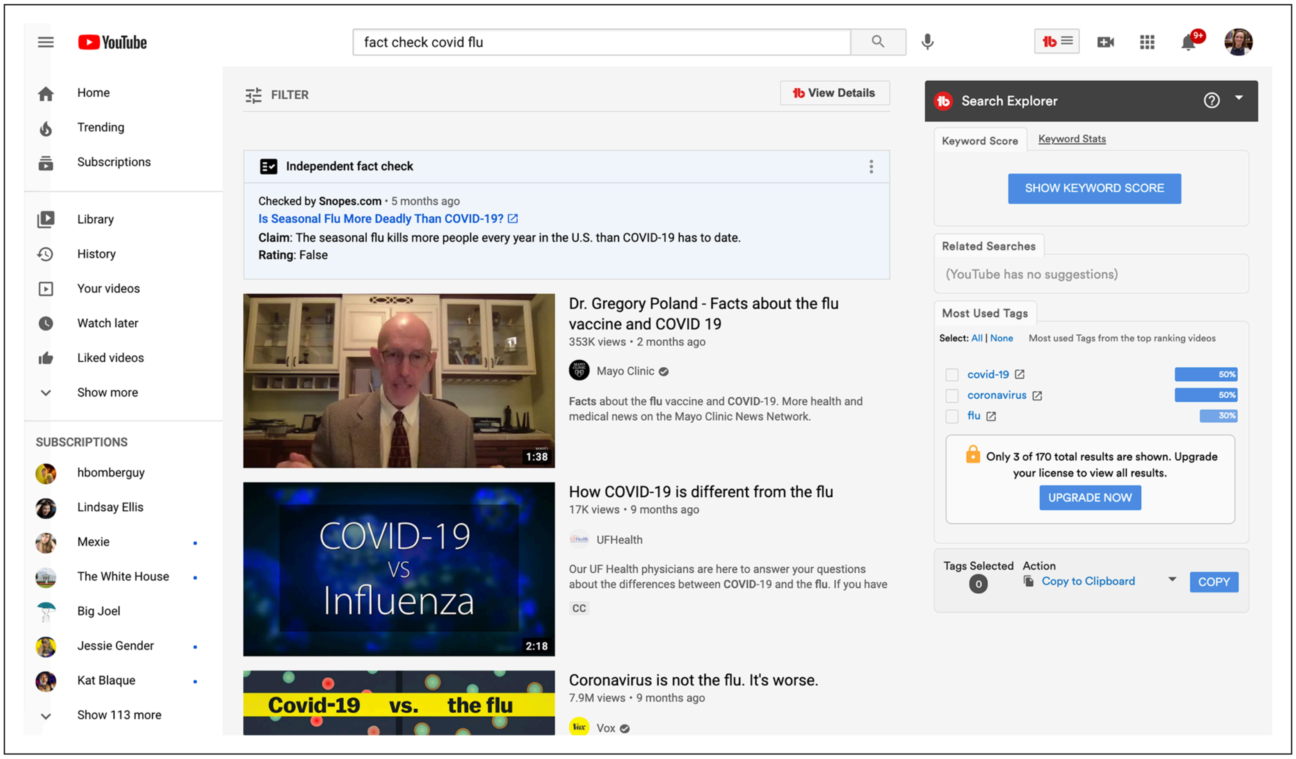
False Information
This post has been reviewed by independent fact-checkers.

Liked by hazeljennings and others
alexraj look at this.
lizunakim this is an insane shot - is it real? 😮
View all 10 comments
Add a comment...
3 HOURS AGO

Figure 4. Facebook's warning label for false information (Facebook, 2019).

buttons on the warning labels compel users to click to learn what it was about the posts' content that earned this label (see Figures 3 and 4).

YouTube did not indicate in any of its communications that it used warning labels in the same way as the other two platforms. Rather than actively label individual videos as



Independent fact check

Checked by Snopes.com • 5 months ago
Is Seasonal Flu More Deadly Than COVID-19?
Claim: The seasonal flu kills more people every year in the U.S. than COVID-19 has to date.
Rating: False

Dr. Gregory Poland - Facts about the flu vaccine and COVID 19
353K views • 2 months ago
Mayo Clinic
Facts about the flu vaccine and COVID-19. More health and medical news on the Mayo Clinic News Network.

How COVID-19 is different from the flu
17K views • 9 months ago
UFHealth
Our UF Health physicians are here to answer your questions about the differences between COVID-19 and the flu. If you have

Coronavirus is not the flu. It's worse.
7.9M views • 9 months ago
Vox

(YouTube has no suggestions)

Most used Tags from the top ranking videos

□	covid-19	50%
□	coronavirus	50%
□	flu	30%

Only 3 of 170 total results are shown. Upgrade your license to view all results.

Tags Selected: 0
Action: Copy to Clipboard

Figure 5. YouTube fact-check information panel.

false or disputed or misleading, YouTube stated that it would display fact-checks in information panels in some search results (see Figure 5):

> There are a few factors that determine whether a fact check information panel will appear for any given search. Most important, there must be a relevant fact check article available from an eligible publisher. And in order to match a viewer's needs with the information we provide, fact checks will only show when people search for a specific claim. For example, if someone searches for "did a tornado hit Los Angeles," they might see a relevant fact check article, but if they search for a more general query like "tornado," they may not. (YT21)

This statement established that users *may* see fact-checks, though did not clarify *when* users could expect to see fact-checks.

In the next section, we explain how the three platforms applied the foregoing policies and practices to the *Plandemic* video, which went viral in early May 2020.

Plandemic as a Case Study of Platform Fact-Checking in Practice

Featuring a discredited scientist who supported claims that "a shadowy cabal of elites is using a global crisis as a cover to profiteer and entrench their power" (Newton, 2020a), *Plandemic* espoused various common conspiracy theories that emerged early on in the COVID-19 crisis. The most prominent false claims the video propagated included that COVID-19 was engineered, not naturally occurring; masks can make people sick; and death counts were being inflated. The video was first published to Facebook, YouTube, and Vimeo on 4 May 2020 (Frenkel, Decker, & Alba, 2020), soon finding audiences on Twitter and Instagram as well (Hatmaker, 2020). It particularly gained steam as people shared it to QAnon, conspiracy theory, and anti-vaccine Facebook groups (Frenkel et al., 2020; Newton, 2020a; Nilsen, 2020), and as microcelebrities, including a celebrity doctor, shared it (Frenkel et al., 2020). In the week that followed the video's debut, it accrued over 8 million views across YouTube, Facebook, Twitter, and Instagram (Frenkel et al., 2020). The "star" of the video gained 130,000 new followers on Twitter following the video's release (Newton, 2020a).

By the time *Plandemic* premiered, the three platforms had already instituted some new policies and practices to address coronavirus misinformation. In response to *Plandemic*, YouTube and Facebook removed the video from their sites on 7 May, 3 days after it was published (Frenkel et al., 2020). Facebook also indicated that they had demoted the video prior to removing it (Newton, 2020a). The decision to remove *Plandemic* hinged on a determination that the specific false claim that wearing a mask can make people sick, could lead to "imminent harm" (Andrews, 2020). According to *The Verge*, YouTube said "it did not recommend 'Plandemic' or surface it 'prominently' in search results" (Newton, 2020a).

Twitter began blocking associated hashtags (e.g., #Plandemic, #PlagueOfCorruption) around the same time that the other two platforms removed the video (Wong & Solsman, 2020). Twitter opted not to remove links to the video's website, instead marking it as "unsafe," which limited its reach, and directed users who clicked on it to a warning message indicating that the content was "potentially spammy or unsafe" (Robertson, 2020).

On 18 August a second, feature-length *Plandemic* film, *Plandemic: Indoctornation*, was released. By this time, Facebook had blocked access to the video's domain, banning users from sharing it (Brewster, 2020). A company spokesman also noted that "This latest video contains COVID-19 claims that our fact-checking partners have repeatedly rated false so we have reduced its distribution and added a warning label showing their findings to anyone who sees it" (Newton, 2020b, p. n.p.). Users who attempted to share the link also received a pop-up message stating the video violated the Community Standards on spam (Zadrozny, 2020). YouTube removed uploads of the full video, but noted that they had not seen many attempts to upload it (Newton, 2020b). YouTube specified that they would deal with clips of the video on a case-by-case basis (Newton, 2020b). Twitter again did not block the video's domain, but instead displayed a warning message to users who clicked, indicating the link was "potentially spammy or unsafe" (Robertson, 2020).

With the benefit of time, the three platforms have managed to vastly minimize access to the *Plandemic* videos. While a handful of stray copies of the videos or clips can still be found on Facebook, Twitter, and YouTube (as evident from keyword searches), they are not easy to stumble upon or even actively seek out. On the one hand, this is a commendable feat: the three platforms effectively scrubbed the videos from their sites. On the other hand, *Plandemic* went viral in the first place *because of* these platforms, and, to some extent, once the first video went viral, significant damage had already been done. It did not matter that the platforms verified the claims of the video through their fact-checking programs, because millions of people around the world learned of *Plandemic* on Facebook, Twitter, and YouTube in the 3 days it took for the platforms to respond. Consistent with what we saw in the platforms' communications, their handling of *Plandemic* suggests a preoccupation with visibility of (certain kinds of) misinformation first and an informed public second. The primary concern was mitigating risk through assessing potential for harm. While fact-checking informed the platforms' decision-making, harm-assessments, rather than false determinations, prompted removal.

Discussion

Our findings suggest that, in the early days of the pandemic, Facebook, Twitter, and YouTube focused their energies primarily on managing the *visibility* of COVID-19 mis/disinformation on their sites via (often vaguely described) policies and practices rife with subjectivity. This represents a rearticulation of the platforms' insistence that they are not arbiters of truth (Gillespie, 2010), as well as an extension of sociotechnical systems they have built to deal with other kinds of problematic content. The three platforms all employed practices for verifying information through fact-checking, relying on a combination of human and algorithmic actors. All three platforms then either demoted, removed, or labeled content deemed false or inaccurate. Although these policies and practices can help mitigate the spread of mis/disinformation, as we saw with *Plandemic*, they do not always effectively or sufficiently limit exposure. To "flatten the curve" of misinformation, as Donovan (2020) wrote, platforms need to become more accountable for the content that they make available to billions of users. As the platforms further integrate fact-checking into the heart of their services, their decisions can shape the broader infrastructure of fact-checking by encouraging solutions that repurpose past practices, but which are vaguely articulated and do not address the full range of factors that contribute to information pollution. Simply put, platforms' focus on limiting the visibility of mis/disinformation may divert attention away from the ideal outcome of fact-checking: a well-informed global public, an outcome that is especially critical during a pandemic. Below we further discuss the ways the platforms policies and practices did not quite accomplish this outcome, and then offer some recommendations.

In our analyses, we saw that the platforms wanted to regulate the visibility of facts and falsities according to subjective (and not openly shared) risk assessments made internally. Recall that both Facebook and Twitter had policies that dictate false information would only be removed if there was a potential for harm, otherwise it would be demoted (Facebook) and labeled (Facebook and Twitter). This creates an essential hierarchy of false COVID-19 information based on judgments about which false or inaccurate claims are likely, as Facebook put it, to "contribute to imminent physical harm" [FB48]. As we saw with *Plandemic*, such judgments are highly subjective: while Facebook identified one claim in the video that they believed could lead to "imminent harm," Twitter seemingly did not judge the video generative of real-world harm, based on its decision not to remove the video. Moreover, this harm-based approach also indicates that the platforms did not believe that COVID-19 mis/disinformation, on its own, could be corrosive. Even mis/disinformation not directly resulting in physical harm can erode trust in institutions, which poses a more pernicious threat to a healthy democracy and citizenry (Kavanagh & Rich, 2018). As we have seen in the spread of mis/disinformation around COVID-19, the circulation of false claims that are not causally related to physical harm have indirectly led to an increase in anti-Asian hate crimes (Reny & Barreto, 2020), allowed the pandemic to surge due to vehement refusal to wear masks and social distance (Hornik et al., 2021), and resulted in attempts to sabotage vaccine distribution (Razek et al., 2021). If platforms are serious about combating mis/disinformation, fact-checking labels and nebulous policies about what constitutes a narrow definition of "harm" likely do little to "flatten the curve" of the infodemic. Instead, platform fact-checking, at least in the early days of the COVID-19 pandemic, seems to be designed to protect the companies against liability for and bad press around real-world harms connected to activity on their sites.

Due to their risk-based approach, Facebook and Twitter also pursued a different aim than fact-checkers generally do. Fact-checking, traditionally, aims to pass a judgment on a claim to determine whether it is true or false. By contrast, by leaving false and inaccurate information up on their sites, though demoting and/or labeling the content, the platforms reaffirmed their underlying "marketplace of ideas" approach to governance (Maddox & Malson, 2020). In other words, the platforms "prioritize[d] free speech and more speech to correct the record" (Maddox & Malson, 2020, p. 8). Although their warning labels may indicate, for example, as Twitter's labels state, that content "conflicts with guidance from public health experts regarding COVID-19," the decision to allow mis/disinformation to remain visible on the sites (though perhaps less visible) implicitly gives some credence to the false claims. This response amounts to an impartial shrug, effectively communicating to users: "Others have determined this is false, but you decide for yourself." When it comes to health mis/disinformation, a "you decide" approach does not make sense. For example, Twitter permitted the *Plandemic* video to remain on the site, explaining that people often "[dispute] the claims in real time with the link included as context" (Brewster, 2020). However, relative to other kinds of mis/disinformation (e.g., in the political realm), medical facts and public health guidelines are fairly clear cut: we judge the veracity of claims based on how well they match the facts and guidance offered by doctors, researchers, and public health experts. When it comes to COVID-19, in most cases, claims that contradict expert medical advice are not matters of "differing opinion." Average individuals are not qualified in general to judge the credibility of medical claims or advice on their own. Moreover, it is questionable whether a user needs to link to false information to dispel it.

As the platforms lean on automated systems to support fact-checking on their sites, they urge a deeper investment in technologies of visibility within the broader fact-checking infrastructure. Certainly, there are many opportunities for automated systems to support the identification of claims to fact-check, as well as the verification and correction of claims (see Graves, 2018). However, developing systems to support fact-checking also requires significant resources (Graves, 2018), which means that the three platforms may have an outsized influence on the design, testing, and application of them. Indeed, Facebook and YouTube's parent company, Google, have been major benefactors of the fact-checking industry in recent years, with Facebook reportedly paying fact-checking organizations hundreds of thousands of dollars each annually (Bell, 2019) and with both platforms creating grant programs to support fact-checking (FB52; YT21). Moreover, the design of such systems shapes the nature and work of fact-checking (Graves, 2018). Our analysis of the platforms' communications indicated that their automated systems were oriented toward regulating the visibility of mis/disinformation via detection and downranking. However, it is not clear whether or how the platforms are developing systems to support the verification of information. Facebook's and YouTube's partnerships with third-party fact-checking organizations have given them room to relieve themselves of responsibility for fact-checking determinations. They may be reluctant to develop automated systems that would support verification, focusing instead on systems that allow them to modulate the relative visibility of "good" and "bad" information. The focus on visibility of content not only creates a loophole for unverified claims to circulate, but also directs the focus toward what is visible to users rather than rethinking the systems that give rise to viral mis/disinformation in the first place.

In short, this article demonstrates how regimes of platform fact-checking embody a cyberlibertarian ideology that drives platform decisions, interpellating their users and fact-checking partners to become subjects to their rules. Platforms, then, urge their users to absorb these ideologies and push fact-checking as a meaningful form of enforcement via their global prominence and their ample revenue streams that allow them to invest in fact-checking around the world. While our data and analysis focused on the U.S. context, the platforms' fact-checking measures extend beyond the U.S. borders. Indeed, the platforms have historically tested measures in other countries prior to implementing them in the United States, although they tend to prioritize the United States in their communications (Karanicolas, 2020). The global impact of platform fact-checking is particularly apparent as different countries have grappled with their own local informational challenges related to the pandemic and elections. Even as new fact-checking organizations emerge in various countries, platforms' investments in fact-checking grant them considerable power in asserting their own values and logics in collective processes of truth-making. As such, we contribute to theoretical discussions across platform and infrastructure studies, journalism studies, science and technology studies, among other fields, about how platforms (re)shape how we make sense of global events like the COVID-19 pandemic, though with a particular Silicon Valley ethos.

Our findings also give insight into some actions platforms could take to shift efforts from the core emphasis on regulating the visibility of mis/disinformation to increasing informedness and healthier, fact-based discourse within their sites. To do this, platforms could devote resources to not only telling people whether something is true or false, but to increasing users' media and digital literacies via interventions within the platform interfaces. Moreover, many of the "informational resources" platforms present to users require a significant level of *health* and *science* literacy, which are exceedingly low in the United States (Salisbury, 2020). As Kearney and colleagues (2020) suggested, investing in "primary preventions" like "pre-bunking" rather than debunking once misinformation has spread are necessary to help users build these literacy skills. Platforms should develop methods of catching mis/disinformation *before* it goes viral. Engaging in modes of *anticipatory* moderation—for example, "virality circuit breakers" for false information (Simpson & Conner, 2020)—will be a critical element in helping to stop the spread of viral misinformation, and platforms must work together and collaborate on these anticipatory actions since viral content moves from platform to platform. As part of this, platforms could devote resources to identifying communities and influencers driving adjacent or related conversations about conspiracy theories. For example, identifying accounts historically responsible for significant proportions of anti-vaccine misinformation can allow for targeted fact-checking during the pandemic that preempts "superspreader" infodemic events as was seen with the *Plandemic* videos. This is especially critical as the United States often acts as a major exporter of online misinformation (Bridgman et al., 2021), and the three platforms investigated in this study have a responsibility, as U.S.-based companies, to prevent this.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kelley Cotter https://orcid.org/0000-0003-1243-0131

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Twitter's communications about this program were not included in our dataset, as we cut off data collection in November 2020. However, we feel obliged to mention it, given its relevance to the study's focus on fact-checking.
2. See https://about.fb.com/news/2019/12/helping-fact-checkers/. This document was included in our full dataset of documents pertaining to fact-checking, but not the narrower COVID-19 dataset.
3. See https://support.google.com/youtube/answer/7554338?hl=en&ref_topic=9387085. This document was included in our full dataset of documents pertaining to fact-checking, but not the narrower COVID-19 dataset.

References

Adair, B., & Stencel, M. (2020). *A lesson in automated journalism: Bring back the humans*. https://www.niemanlab.org/2020/07/a-lesson-in-automated-journalism-bring-back-the-humans/

Amazeen, M. A., Thorson, E., Muddiman, A., & Graves, L. (2015). *A comparison of correction formats: The effectiveness and effects of rating scale versus contextual corrections on misinformation*. American Press Institute. https://www.americanpressinstitute.org/wp-content/uploads/2015/04/The-Effectiveness-of-Rating-Scales.pdf

Ananny, M. (2018). *The Partnership Press: Lessons for platform-publisher collaborations as Facebook and news outlets team to fight misinformation*. Tow Center for Digital Journalism. https://www.cjr.org/tow_center_reports/partnership-press-facebook-news-outlets-team-fight-misinformation.php

Andrews, T. M. (2020, May 7). Facebook and other companies are removing viral "Plandemic" conspiracy video. *The Washington Post*. https://www.washingtonpost.com/technology/2020/05/07/plandemic-youtube-facebook-vimeo-remove/

Ball, P., & Maxmen, A. (2020). The epic battle against coronavirus misinformation and conspiracy theories. *Nature*, *581*(7809), 371–374. https://doi.org/10.1038/d41586-020-01452-z

Bell, E. (2019). *The fact-check industry*. https://www.cjr.org/special_report/fact-check-industry-twitter.php/

Brewster, J. (2020, August 18). Facebook bans users from sharing debunked "Plandemic" movie. *Forbes*. https://www.forbes.com/sites/jackbrewster/2020/08/18/facebook-bans-users-from-sharing-debunked-plandemic-movie/

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., & Zhilin, O. (2020). The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-028

Bridgman, A., Merkley, E., Zhilin, O., Loewen, P. J., Owen, T., & Ruths, D. (2021). Infodemic pathways: Evaluating the role that traditional and social media play in cross-national information transfer. *Frontiers in Political Science*, *3*, Article 648646. https://doi.org/10.3389/fpos.2021.648646

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2), 1–13. https://doi.org/10.1177/2056305120936636

Coleman, K. (2021, January 25). *Introducing Birdwatch, a community-based approach to misinformation*. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html

Donovan, J. (2020). Social-media companies must flatten the curve of misinformation. *Nature*. https://doi.org/10.1038/d41586-020-01107-z

Facebook. (2019, October 21). *Helping to protect the 2020 US elections*. https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/

Ferrara, E., Cresci, S., & Luceri, L. (2020). Misinformation, manipulation, and abuse on social media in the era of COVID-19. *Journal of Computational Social Science*, *3*(2), 271–277. https://doi.org/10.1007/s42001-020-00094-5

Fischer, S. (2020, October 13). Fact-checking goes mainstream in Trump era. *Axios*. https://www.axios.com/fact-checking-trump-media-baad50cc-a13f-4b73-a52a-4cd9e63bd2fc.html

Frenkel, S., Decker, B., & Alba, D. (2020, May 20). How the "Plandemic" movie and its falsehoods spread widely online. *The New York Times*. https://www.nytimes.com/2020/05/20/technology/plandemic-movie-youtube-facebook-coronavirus.html

Garrett, R. K., Weeks, B. E., & Neo, R. L. (2016). Driving a wedge between evidence and beliefs: How online ideological news exposure promotes political misperceptions. *Journal of Computer-Mediated Communication*, *21*(5), 331–348. https://doi.org/10.1111/jcc4.12164

Ghebreyesus, T. A. (2020). *Munich Security Conference*. https://www.who.int/director-general/speeches/detail/munich-security-conference

Gillespie, T. (2010). The politics of "platforms." *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

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

Gillespie, T. (2018b). Regulation of and by platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE handbook of social media* (pp. 254–278). SAGE. https://doi.org/10.4135/9781473984066.n15

Gillespie, T. (2020). Platforms throw content moderation at every problem. In M. Zimdars & K. McLeod (Eds.), *Fake news: Understanding media and misinformation in the digital age* (pp. 329–339). The MIT Press.

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Graves, L. (2018, February 28). *Understanding the promise and limits of automated fact-checking*. https://www.digitalnewsreport.org/publications/2018/factsheet-understanding-promise-limits-automated-fact-checking/

Graves, L., & Amazeen, M. A. (2019). Fact-checking as idea and practice in journalism. In L. Graves & M. A. Amazeen (Eds.), *Oxford research encyclopedia of communication*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.808

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology*, *17*(1), 42–108.

Gyenes, N., & Mina, A. X. (2018, August 30). How misinfodemics spread disease. *The Atlantic*. https://www.theatlantic.com/technology/archive/2018/08/how-misinfodemics-spread-disease/568921/

Hatmaker. (2020, May 7). Platforms scramble as "Plandemic" conspiracy video spreads misinformation like wildfire. *TechCrunch*. https://techcrunch.com/2020/05/07/plandemic-video-judy-mikovits/?guccounter=1

Hornik, R., Kikut, A., Jesch, E., Woko, C., Siegel, L., & Kim, K. (2021). Association of COVID-19 misinformation with face mask wearing and social distancing in a nationally representative us sample. *Health Communication*, *36*(1), 6–14. https://doi.org/10.1080/10410236.2020.1847437

Horwitz, J. (2021, September 13). Facebook says its rules apply to all. Company documents reveal a secret elite that's exempt. *Wall Street Journal*. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353

Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to online disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics*, *25*(3), 493–516. https://doi.org/10.1177/1940161219900126

IFCN. (n.d.). *IFCN code of principles*. https://www.ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles

Kaiser, J., Rauchfleisch, A., & Córdova, Y. (2021). Fighting Zika with honey: An analysis of YouTube's video recommendations on Brazilian YouTube. *International Journal of Communication*, *15*, 1244–1262.

Karanicolas, M. (2020, November 16). The countries where democracy is most fragile are test subjects for platforms' content moderation policies. *Slate*. https://slate.com/technology/2020/11/global-south-facebook-misinformation-content-moderation-policies.html

Kavanagh, J., & Rich, M. D. (2018). *Truth decay: An initial exploration of the diminishing role of facts and analysis in American public life*. RAND Corp. https://doi.org/10.7249/RR2314

Kearney, M. D., Chiang, S. C., & Massey, P. M. (2020). The Twitter origins and evolution of the COVID-19 "plandemic" conspiracy theory. *Harvard Kennedy School Misinformation Review*, *1*(3), 1–18. https://doi.org/10.37016/mr-2020-42

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Maddox, J., & Malson, J. (2020). Guidelines without lines, communities without borders: The marketplace of ideas and digital manifest destiny in social media platform policies. *Social Media + Society*, *6*(2), 1–10. https://doi.org/10.1177/2056305120926622

Marietta, M., Barker, D. C., & Bowser, T. (2015). Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum*, *13*(4), 577–596. https://doi.org/10.1515/for-2015-0040

Newton, C. (2020a, May 12). *How the "Plandemic" video hoax went viral*. https://www.theverge.com/2020/5/12/21254184/how-plandemic-went-viral-facebook-youtube

Newton, C. (2020b, August 19). Platforms successfully stopped a COVID conspiracy video from going viral. *The Verge*. https://www.theverge.com/interface/2020/8/19/21373820/plandemic-indoctornation-facebook-youtube-twitter-removal-block-covid-hoax-block

Nilsen, J. (2020, October 7). *Distributed amplification: The Plandemic documentary*. https://mediamanipulation.org/case-studies/distributed-amplification-plandemic-documentary

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, *42*(3), 939–960. https://doi.org/10.1007/s11109-019-09528-x

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical Care*, *51*(2), 127–132. https://doi.org/10.1097/MLR.0b013e318279486b

Pasternack, A. (2020, August 20). *How Facebook quietly pressures its independent fact-checkers*. https://www.fastcompany.com/90538655/facebook-is-quietly-pressuring-its-independent-fact-checkers-to-change-their-rulings

Porter, E., & Wood, T. J. (2020). *Misinformation on the Facebook News Feed: Experimental Evidence* [Preprint]. https://doi.org/10.31219/osf.io/r3mvw

Razek, R., Alsup, D., & Waldrop, T. (2021). *Wisconsin pharmacist who left Covid-19 vaccine out pleads guilty*. https://www.cnn.com/2021/01/26/us/wisconsin-pharmacist-covid-vaccine-guilty-plea/index.html

Reny, T. T., & Barreto, M. A. (2020). Xenophobia in the time of pandemic: Othering, anti-Asian attitudes, and COVID-19. *Politics, Groups, and Identities*. Advance online publication. https://doi.org/10.1080/21565503.2020.1769693

Roberts, S. T. (2018). Digital detritus: "Error" and the logic of opacity in social media content moderation. *First Monday*. https://doi.org/10.5210/fm.v23i3.8283

Robertson, A. (2020, August 18). Facebook blocks users from linking to new Plandemic hoax video. *The Verge*. https://www.theverge.com/2020/8/18/21374081/plandemic-indoctornation-conspiracy-video-facebook-misinformation

Salisbury, M. (2020, July 29). How America's low science literacy fueled the COVID crisis. *Techonomy*. https://techonomy.com/2020/07/science-literacy-and-americas-covid-crisis/

Simpson, E., & Conner, A. (2020). *Fighting coronavirus misinformation and disinformation*. Center for American Progress. https://www.americanprogress.org/issues/technology-policy/reports/2020/08/18/488714/fighting-coronavirus-misinformation-disinformation/

Smith, R., Wardle, C., & Cubbon, S. (2020). *Under the surface: Covid-19 vaccine narratives, misinformation and data deficits on social media*. First Draft. https://firstdraftnews.org/vaccine-narratives-report-summary- november-2020

Stencel, M., & Luther, J. (2021). *Fact-checking census archives*. https://reporterslab.org/tag/fact-checking-census/

Stewart, E. (2021). Detecting fake news: Two problems for content moderation. *Philosophy & Technology*, *34*, 923–940. https://doi.org/10.1007/s13347-021-00442-x

Suárez, E. (2020, March 31). *How fact-checkers are fighting coronavirus misinformation worldwide*. https://reutersinstitute.politics.ox.ac.uk/risj-review/how-fact-checkers-are-fighting-coronavirus-misinformation-worldwide

Swire-Thompson, B., & Lazer, D. (2020). Public health and online misinformation: Challenges and recommendations. *Annual Review of Public Health*, *41*(1), 433–451. https://doi.org/10.1146/annurev-publhealth-040119-094127

Tan, A. S. L., Lee, C., & Chae, J. (2015). Exposure to health (mis) information: Lagged effects on young adults' health behaviors and potential pathways. *Journal of Communication*, *65*(4), 674–698. https://doi.org/10.1111/jcom.12163

Twitter. (n.d.). *Birdwatch guide: Overview*. https://twitter.github.io/birdwatch/about/overview/

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society*. Oxford University Press.

Villasenor, J. (2020, November 23). *How to deal with AI-enabled disinformation*. https://www.brookings.edu/research/how-to-deal-with-ai-enabled-disinformation/

Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, *37*(1), 136–144. https://doi.org/10.1080/10584609.2020.1716500

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

Wong, Q., & Solsman, J. E. (2020, May 8). Facebook, YouTube and Twitter struggle with viral Plandemic conspiracy video. *CNET*. https://www.cnet.com/news/facebook-youtube-twitter-viral-plandemic-conspiracy-video/

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., & Patil, S. (2020, April 25–30). *Effects of credibility indicators on social media news sharing intent* [Conference session]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI. https://doi.org/10.1145/3313831.3376213

Zadrozny, B. (2020, August 18). *Facebook not allowing users to post the new plandemic link* [Tweet]. https://twitter.com/BrandyZadrozny/status/1295781499556499456

Author Biographies

Kelley Cotter (PhD, Michigan State University) is an Assistant Professor in the College of Information Sciences and Technology at the Pennsylvania State University. Her research interests include the social impacts of algorithms and big data; algorithmic, digital, and media literacies; and platform governance.

Julia R. DeCook (PhD, Michigan State University) is an Assistant Professor of Advocacy and Social Change in the School of Communication at Loyola University Chicago. Her research interests include platform governance, online hate, race and gender, digital culture, and social justice and technology.

Shaheen Kanthawala (PhD, Michigan State University) is an Assistant Professor in the Department of Journalism and Creative Media at the University of Alabama. Her research interests include health information and technologies, emerging technologies, and platform governance.