

Look Who's Watching: Platform Labels and User Engagement on State-Backed Media Outlets

American Behavioral Scientist

2024, Vol. 68(10) 1325–1344

© 2023 SAGE Publications

Article reuse guidelines:



sagepub.com/journals-permissions

DOI: 10.1177/00027642231175639

journals.sagepub.com/home/abs



**Samantha Bradshaw¹, Mona Elswah²,
and Antonella Perini²** 

Abstract

Recently, social media platforms have introduced several measures to counter misleading information. Among these measures are “state-media labels” which help users identify and evaluate the credibility of state-backed news. YouTube was the first platform to introduce labels that provide information about state-backed news channels. While previous work has examined the efficiency of information labels in controlled lab settings, few studies have examined how state-media labels affect users’ perceptions of content from state-backed outlets. This article proposes new methodological and theoretical approaches to investigate the effect of state-media labels on users’ engagement with content. Drawing on a content analysis of 8,071 YouTube comments posted before and after the labeling of five state-funded channels (Al Jazeera English [AJE], China Global Television Network, Russia Today [RT], TRT World, and Voice of America [VOA] News), this article analyses the effect that YouTube’s labels had on users’ engagement with state-backed media content. We found that the labels had no impact on the amount of likes that the videos received before and after the policy introduction, except for RT which received less likes after it was labeled. However, for RT, comments left by users were associated with 30% decrease in the likelihood of observing a critical comment following the policy implementation, and a 70% decrease in likelihood of observing a critical comment about RT as a media source. While other state-funded broadcasters, like AJE and VOA News, received fewer critical comments after YouTube introduced its policy; this relationship was associated with how political the video was, rather than the

¹Stanford Internet Observatory, Stanford University, Stanford, CA, USA

²Oxford Internet Institute, Oxford University, Oxford, UK

* Samantha Bradshaw is also affiliated to School of International Service, American University, Washington DC, USA

Corresponding Author:

Samantha Bradshaw, Stanford Internet Observatory, Freeman Spogli Institute for International Studies, Stanford University, Encina Hall, Stanford, CA 94305-6055, USA.

Email: sam.r.bradshaw@gmail.com

policy change. Our study contributes to the ongoing discussion on the efficacy of platform governance in relation to state-backed media, showing that audience preferences impact the effectiveness of labels.

Keywords

state media, platform governance, selective exposure, YouTube, labeling, perceptions, credibility

Introduction

The spread of disinformation online has become a critical question for political communication research and policymaking more broadly (Allcott & Gentzkow, 2017; Freelon & Wells, 2020; Ha et al., 2021; Lazer et al., 2018; Weeks & Gil de Zúñiga, 2021). To combat its spread, social media platforms have introduced a range of policy responses (Taylor et al., 2018). Information labeling is one mechanism that platforms have adopted to provide users with more information about the credibility of online news. In February 2018, YouTube was the first platform to implement a *state-funding* label for U.S. audiences, providing notices below videos uploaded by news broadcasters that receive some level of government or public funding (Samek, 2018). In addition to state-funded labels, YouTube also provides *credibility* labels below scientific or politicized topics that have been subject to misinformation, linking to third-party sources and additional fact-checked information (Mohan & Kyncl, 2018).

YouTube is not the only platform to apply *source* and *credibility* labels to news and information online. In 2016, Facebook introduced a credibility policy that labeled news stories containing false information with a “disputed” flag (Mosseri, 2017). This policy eventually replaced the “disputed” flag with a “related articles” notification, showing users fact-checked information below stories that contained false information (Lyons, 2017). In June 2020, Facebook also implemented *source* labels for state-funded outlets that were “wholly or partially under the editorial control of their government” (Gleicher, 2020). Twitter began rolling out *credibility* labels in 2019 and 2020, including notices under tweets by politicians or world leaders that breach Twitter’s rules (Twitter Safety, 2019), as well as notices under “synthetic and manipulated media” such as deep fake videos (Roth & Achuthan, 2020), and coronavirus mis- and disinformation (Roth & Pickles, 2020). In August 2020, Twitter implemented a *source* labeling policy for state-affiliated media accounts from countries in the United Nations Security Council—Russia, China, the United States, the United Kingdom, and France—however, in practice it only labeled Russian and Chinese outlets (Ha, 2020).

Labeling information is intended to make users critical readers of news by providing them with additional context about the information’s source and accuracy. These labels aim to inoculate users by exposing them to weakened, refutational, or corrected arguments, which can help refute malicious or conspiratorial information (Cook et al., 2017; Lim & Ki, 2007; Van der Linden et al., 2017; Nassetta & Gross, 2020; Pfau

et al., 1997). While there are only a few studies examining the impact of labels introduced by platforms, early evidence suggests mixed results. Warning labels can reduce the sharing of mis-and-disinformation (Mena, 2020), but if users have already been exposed to false claims, seeing a “disputed” flag after the fact will not alter their original belief (Pennycook et al., 2018). Labeling information can also lead to unintended consequences, such as the “implied truth effects,” where unlabeled false claims are more likely to be interpreted as true (Pennycook et al., 2020). And emphasizing publisher information can increase the chance that users will view truthful claims as false (Dias et al., 2020). While these studies shed light on the impact and effect of *correction* labels on users’ perceptions of news, few systematic studies have examined the effect of *source* labels, particularly in the context of state-backed media outlets. This is important given that state-backed media outlets—especially those from authoritarian regimes—have come under increasing pressure for spreading propaganda online (Diamond et al., 2016; Elswah & Howard, 2020; McFaul & Kass, 2019; Molter, 2020).

Studies related to selective exposure suggest that people tend to consume news that aligns with their ideology, values, and preexisting beliefs (Chaffee & Miyo, 1983; Hart et al., 2009; Masip et al., 2018). Through this lens, one reason users select information that confirms their worldview is to avoid the cognitive dissonance of exposure to radically new or challenging information (Davis & Jurgenson, 2014; Festinger, 1957). Indeed, studies of selective exposure show that users are more likely to engage with social media content that is in line with their beliefs and opinions (Cinelli et al., 2020; Masip et al., 2018; Skoric et al., 2018), and challenge, remove, or ignore conflicting information to reduce dissonance (Bai et al., 2019). Because social media labels can present ideas that conflict or challenge users’ worldviews, we ask: when do state-funded labels affect users’ perceptions and engagement with content from state-funded media outlets on YouTube?

The aim of this study is to understand the efficacy of state-media labels on users’ perception of and engagement with state-backed media content online. First, we aim to understand under what conditions state-media labels inoculate users against potential manipulative or overt propaganda campaigns from state-backed media sources. Second, state-backed media are diverse in their content, coverage, and editorialization. However, when YouTube implemented its labeling policy, it received significant criticism for including editorially independent public broadcasters—like the BBC or PBS News—alongside state-backed propaganda outlets (Radsch, 2018; Shaban, 2018). Thus, we also aim to understand comparatively how state-media labels might affect the audiences of different state-backed channels.

Our findings contribute to the ongoing academic discussions about the efficacy of information labels on user perceptions by highlighting the importance of the audience. Our work compliments the growing body of empirical research that demonstrates the unintended consequences of labeling, strengthening the field’s overall argument that labeling is not the ultimate solution to mis-and-disinformation. While previous experiments highlight the importance of timing or effect that labels have on a user’s ability to determine facts, by examining user-generated content we demonstrate that platforms should not expect homogenous effects of labels on their users, and that online users have varying ideologies and political leanings that should be considered when introducing misinformation countermeasures.

Literature Review

Engagement, Exposure, and Cognitive Dissonance

The rich affordances of social media platforms allow users to interact with social media platforms in multiple ways, including liking or disliking, following, commenting, uploading, or sharing content online (Khan, 2017). However, different motivations drive different types of interaction with online content (Kim & Yang, 2017; Swani & Labrecque, 2020). Commenting on social media content is considered a higher act of public expression compared to watching, reading, or liking content, which are considered lower level acts of engagement (Aldous et al., 2019; Muntinga et al., 2011). While engagement with social media content is considered a key indicator of the content popularity and the success of the content creator (Schreiner et al., 2019), it can also represent political information about users and audiences because of how users “selectively expose” themselves to content that is consistent with their worldview.

There is a rich corpus of literature examining how individuals play a role in exercising information preferences in both the online and offline world. In the offline world, research on selective exposure shows that people tend to select traditional media and broadcasting sources they want to consume, and choose to associate with certain political groups, including parties, community associations, or candidates (Chaffee & Miyo, 1983). The more an individual is interested in the subject the more likely they are to pay selective attention to sources and ideas that fit within their worldview (Berelson & Steiner, 1964). Many of the studies on selective exposure are being updated to account for the affordances of social media platforms, showing that users are more likely to consume and interact with social media content that supports, rather than contradicts, their worldview (Cinelli et al., 2020; Masip et al., 2018; Quattrocioni et al., 2016). In addition to exposing themselves to content selectively by following, watching or reading sources of news online, users are also more likely to *engage* with content that they perceive as believable and in line with their established beliefs (Kim & Dennis, 2019). This includes commenting, liking, and sharing information that conforms to their political preferences.

One explanation for selective exposure is that users do not want to face the cognitive dissonance of exposure to new or challenging information (Chaffee & Miyo, 1983; Cotton, 1985; Cotton & Hieser, 1980). Based on the Cognitive Dissonance Theory, people tend to selectively consume information that is in consistence with their cognition to avoid discomfort of holding two contradictory thoughts simultaneously (Festinger, 1957; Jeong et al., 2019). The decision to read a certain online news story is motivated by the need for opinion reinforcement and an aversion to ideas that challenge one's perspectives (Garrett, 2009). In the context of social media, users follow and share content that supports their beliefs, and eschew information that contradict their beliefs, known as *selective avoidance* (An et al., 2013; Garrett, 2009; Johnson & Kaye, 2013). Because of context collapse, users have expressed very real and jarring experiences when presented with unexpected information on social media (Davis & Jurgenson, 2014). This negative experience motivates users to selectively consume certain information, while also selectively avoiding other online content and interactions to reduce the dissonance of conflicting information (Bai et al., 2019; Bradshaw & Howard, 2018).

Cognitive Dissonance, Belief Perseverance, and News Corrections

People avoid discomfoting information in general. However, when people are confronted with additional contradictory evidence, they might adopt a dissonance reduction strategy, depending on the magnitude and importance of the discrepancy (Festinger, 1957; Hardyck & Kardush, 1968; Jost, 2015; Vaidis & Bran, 2018). Belief perseverance, or the idea that individuals maintain their beliefs despite new information that refutes its accuracy, is one way that individuals attempt to avoid cognitive dissonance (Anderson et al., 1980; Anglin, 2019; Cohen et al., 2000; Nestler, 2010). In some studies, on belief perseverance, presenting individuals with additional facts does not always lead to a refutation, even they are based on a weak or inconclusive evidence (Anderson et al., 1980).

When it comes to news corrections, many experiments have consistently shown that corrections have little to no effect on a person's perception of truth (Thorson, 2016). Studies have shown that even when presented with correct information, false beliefs continue to persist (Cobb et al., 2013; Ecker et al., 2011; Thorson, 2016). In some cases, exposure to corrections can even "backfire," with the belief in inaccurate information becoming even stronger after being presented with a correction (Nyhan & Reifler, 2010). While concerns about the "backfire" effect continue, several studies have failed to find or replicate this effect (e.g., Nyhan et al., 2019; Schmid & Betsch, 2019; Wood & Porter, 2019). However, when subsequent backfires have been found, they have been mostly identified in small political or attitudinal subgroups (Swire-Thompson et al., 2020). Indeed, ideological cues have been shown to influence partisan perceptions of news credibility (Garrett, 2009; Nyhan & Reifler, 2010).

Nevertheless, much of the recent literature suggests that audiences will accept authoritative cues over ideological ones and update their beliefs in the direction of the fact check (Bullock et al., 2013; Duncan, 2020; Wood & Porter, 2019). In the context of social media labels, some experimental research has demonstrated that being exposed to a "disputed flag" limits the likelihood of that user sharing information with a warning label (Mena, 2020; Pennycook et al., 2020). Studies that focus on the effect of disputed flags have found a modest reduction in the belief of fake news stories (Gao et al., 2018; Pennycook et al., 2020). In particular, when disputed flags reduced the effect of mis-and-disinformation, users were exposed to the flag *before* the false or misleading information (Pennycook et al., 2018). Similarly, Nassetta and Gross (2020) found that labels could mitigate the influence of misinformation if they were *noticed* by the users, demonstrating the importance of placement in addition to timing.

Despite the positive impact that labels can have on inoculating users against mis-and-disinformation, scholars have also identified some unintended consequences of labeling practices. Unlabeled fake news stories were more likely to be perceived as true, a phenomenon called the "implied truth effect" (Pennycook et al., 2020). Others found that exposure to a "disputed" or "rated false" flag had no effect on the perceived accuracy of unlabeled false headlines but instead led to a decreased belief in the accuracy of true headlines (Clayton et al., 2019). When it came to *source* labels, increasing the visibility of publishers was "ineffective and perhaps even counterproductive," and

showing participants headlines with publisher information had no significant effect on whether participants believed the information to be accurate (Dias et al., 2020). Overall, research on the impact of news corrections and information labels presents inconclusive results, mainly from lab-controlled experiments. In this article, we explore the effect of information labels from the users' point of view, evaluating user comments before and after the implementation of YouTube's state-funding labeling policy in 2018.

Methodology

Sampling Strategy

Five state-funded news channels were selected for this study: Al Jazeera English (AJE), China Global Television Network (CGTN), RT (formerly Russia Today), TRT World, and Voice of America (VOA) News (see Table 1 and see Supplemental File for more background information). We selected these channels for three reasons: first, despite being a transparency initiative, YouTube's labeling policy did not list which state-funded media outlets would be labeled. Several journalistic investigations at the time found that small or obscure state-funded channels were not accurately labeled (Kofman, 2019), and YouTube also stated in its announcement that labeling would be an ongoing process (Samek, 2018). Thus, major state-funded news outlets were selected to ensure that the labels were accurately added to the videos uploaded by prominent state-funded channels.¹ Second, all five channels target U.S. citizens with online content. Since YouTube's state-funded notice policy was not implemented globally and instead rolled out to an U.S. audience starting on February 2, 2018, it was important to select channels that would have a large U.S. audience who would notice the flag. Third, despite being state funded, all five channels are different in terms of their organizational culture, editorial policies, and practices, as well as geopolitical aims. Some state-funded news providers, like VOA and AJE, have criticized YouTube's notice policy for treating sources that provided independent journalism as comparable to known propaganda outlets like RT or Sputnik. By looking comparatively across a range of state-funded outlets, this article can provide an insight into whether these concerns are justified.

Data Collection

In this study, we focus on user-generated YouTube comments to understand the efficacy of the state-media labels. After selecting the five channels, we used a python script to query the YouTube API for a list of videos uploaded by the five outlets between January 30, 2018 and February 1, 2018, as well as February 3, 2018 and February 5, 2018. These dates represent 3 days before and 3 days after the notice implementation. There is no information on what time on February 2, 2018 YouTube implemented the notice policy. To ensure videos (and their subsequent comments) could be evaluated before and after the policy implementation, February 2, 2018 was excluded from the video and comment sample. In addition to the video titles and date

Table 1. A Summary of the Information Available on the Five Channels.

Channel	Launching date	Country	Evidence on government interference	Funding model	Social media labels	Annual budget in US\$
AJE	2006	Qatar	NA	State funded	YouTube	NA
CGTN	2000	China	Available	State funded	YouTube, Twitter, and Facebook	NA
RT	2005	Russia	Available	State funded	YouTube, Twitter, and Facebook	400 million
TRT World	2015	Turkey	Available	Public broadcaster	YouTube	77–155 million
VOA News	1942	USA	NA	State funded	YouTube	218.5 million

Source. Based on the available literature on these channels. An overview of these channels can be found in the Supplemental File.

Note. Government interference was identified based on the literature and articles. NA refers to the absence of available information related to this subject. AJE=Al Jazeera English; CGTN=China Global Television Network; RT=Russia Today; VOA=Voice of America.

of upload, we collected data about the video tags, number of views, number of video likes, number of video dislikes, and number of comments. Table 2 summarizes the number of YouTube videos each channel uploaded during the study timeframe and provides summary statistics about engagement.

From this large sample of videos, we selected the top-three most commented on videos for each day in the time frame. We selected a small sample of videos that were most widely discussed by users to ensure that the content was watched and consumed. In total, we selected 88 videos uploaded by the five state-funded media outlets on YouTube. These videos covered a range of political and economic news, including Trump’s State of the Union Address, the downing of a Russian fighter pilot in Syria, child labor in China, and stock market forecasts.

Once we built a sample of videos, we queried the YouTube API for all the user comments associated with them. Because all the videos in the sample were still active videos at the time of data collection, users continued to watch and leave comments about this content. Since this article is concerned with the impact that the state-sponsored label had on users’ comments and perceptions of these sources, we only selected comments left by users up to 1 month after the video was posted (until the end of February 2018). In total, 8,071 comments from 88 videos were coded for this study.

Coding and Typology Building: Assessing User Criticality

To determine how users responded to content from state-funded media organizations, we developed a typology using a grounded theory of coding (Charmaz, 2006). We developed a list of 12 discourse types that described the ways users actively engaged in discussion about content from the state-funded channels. High-level sentiment scores were created to evaluate whether users (a) supported the channel or ideas

Table 2. Videos, Views, (Dis)Likes, and Comments (January 30–February 1 and February 3–February 5).

Channel/state-funded media	Number of videos uploaded	Number of video views	Number of video likes	Number of video dislikes	Number of comments
AJE	52	1,403,493	11,220	13,587	5,977
CGTN	60	2,791,087	8,871	995	2,682
RT	57	2,471,910	24,460	6,256	18,150
TRT World	60	1,063,819	7,139	1,519	2,097
VOA News	47	413,323	1,834	1,214	2,200

Source. Authors' analysis of videos collected through the YouTube API.
Note. AJE = Al Jazeera English; CGTN = China Global Television Network; RT = Russia Today; VOA = Voice of America.

presented in the video; (b) criticized the channel or ideas presented in the video; (c) neither supported nor criticized the channel or ideas presented in the video. The full codebook with descriptions can be found in the Supplemental Appendix for this article.

Three coders were trained on the typology through an interactive process where comments would be coded individually, decisions would be compared, disagreements would be discussed, and a final decision would be made. Coders individually coded 25% (or 2,021 comments) to measure intercoder reliability. Krippendorff's Alpha for high-level code categories (supportive, critical, or neutral) was $\alpha = .87$, suggesting substantial agreement between the coders. Krippendorff's Alpha for the subcategories was $\alpha = .81$. This decrease in agreement was expected, given the fuzzy nature of political communication and determining the meaning from the written comments. With an agreeable intercoder reliability established, the remaining 6,051 comments were split between the three coders who individually coded the remaining sample of comments. In addition to coding comments, the coders also coded videos they watched for the level of politicization. This was done on an ordinal scale of not political, political, and highly political. Krippendorff's Alpha for the video scale was $\alpha = .88$, suggesting strong agreement between coders. All disagreements for both videos and comments were discussed by the authors and a final decision was made.

Analysis

State-Funded Flags and User Engagement

First, to evaluate the impact of state-funded flags on user engagement with state-funded media channels, we collected data about the number of likes and dislikes that the videos received. We calculated the average number of likes and dislikes per video view, before and after the policy change (see Table 3). For all five channels, there was an overall decline in the number of likes that the videos received after the policy change, proportional to the views it received. We conducted a Welch Two-Sample *T*-Test to examine the statistical significance of the difference in means between the

Table 3. Welch Two-Sample T-Test. Comparison of Likes and Dislikes Before and After YouTube State-Funded Policy Implementation.

Post Engagement	Full dataset	AJE	CGTN	RT	TRT World	VOA News
Likes (per 1,000 views)						
Before (n)	12.4	10.4	1.8	18.4	16.6	5
After (n)	4.1	8.7	1.1	7.1	9.2	3.3
Dislikes (per 1,000 views)						
Before (n)	1.4	0.75	0.25	3.0	1.7	2.8
After (n)	0.7	0.14	0.14	0.09	1.1	2.1
Likes						
T-statistic (p value)	-0.51 (0.60)	-0.09 (0.92)	-0.70 (0.49)	-1.99 (0.006)*	-0.43 (0.66)	-0.24 (0.80)
Dislikes						
T-statistic (p value)	0.73 (0.46)	-0.12 (0.90)	1.63 (0.14)	-0.13 (0.89)	0.89 (0.39)	-1.42 (0.17)

Note. AJE= Al Jazeera English; CGTN= China Global Television Network; RT= Russia Today; VOA News =Voice of America News.

*Indicates the coefficient estimate is statistically significant at the 99% confidence level.

average of likes and dislikes, before and after the policy implementation (see Table 3). While there was a decline in the number of likes received, the average in means before and after the policy change was only statistically significant for RT. There were no statistically significant relationships between the two averages of dislikes, before and after the policy implementation for any of the channels.

State-Funded Flags and User Comments

How did the introduction of the YouTube flag influence user comments and how critical they were of the content coming from these channels? We first compared the proportions of critical, supportive, and neutral comments before and after the policy change. Following the introduction of YouTube’s state-funded labeling policy, user comments became less critical of content from AJE (34.6–25.4%) and RT (19.6–16.1%). At the same time, these channels showed an increase in the proportion of supportive comments after the policy change. AJE had the largest (26.8–41.9%) increase in the proportion of supportive comments after the policy change, while analysis of RT showed only a slight increase in supportive comments. In contrast, TRT World and CGTN received a larger proportion of critical comments after the implementation of YouTube’s state-funded labeling policy. CGTN only saw a small (25.7–28.0%) increase in critical comments after the policy change, while TRT World had a very large (27.9–44.8%) increase in the proportion of critical comments. We conducted a Pearson’s chi-squared test to assess the dependence between YouTube’s policy change and the types of comments posted to videos and found a statistically significant relationship across the entire dataset (p value=1.26e–07). The test statistic was also statistically significant when examining the dependence of these variables for AJE, RT, and TRT World (see Table 4).

Table 4. Proportion of Comment Sentiment Before and After YouTube State-Funded Policy Implementation, by State-Funded Channel.

Content Sentiment	Full dataset	AJE	CGTN	RT	TRT World	VOA News
Supportive comments						
Before (%)	11.5%	26.8%	44.3%	55.6%	45.3%	30.1%
After (%)	36.1%	41.9%	39.8%	56.1%	23.5%	33.7%
Critical comments						
Before (%)	7.1%	34.6%	25.7%	19.6%	27.9%	34.2%
After (%)	15.7%	25.4%	28.0%	16.1%	44.8%	34.7%
Neutral comments						
Before (%)	8.1%	38.9%	30.0%	24.8%	26.7%	35.1%
After (%)	21.6%	32.7%	32.3%	27.9%	31.7%	31.5%
Chi squared						
x-Squared	31.6	34.12	0.41	9.27	17.16	1.6
p Value	1.26e-07*	3.91e-08*	0.81	.009*	.0002*	0.4

Note. AJE= Al Jazeera English; CGTN= China Global Television Network; RT= Russia Today; VOA News= Voice of America News.
*Indicates the coefficient estimate is statistically significant at the 99% confidence level 3.

We then performed a logit regression analysis asking: (a) does the YouTube policy change have an impact on the likelihood that a comment about content will be critical (see Table 5) and (b) does the YouTube policy change have an impact on the likelihood that a comment about the source will be critical (see Table 6)? We found that YouTube’s policy change was associated with a lower likelihood of getting a critical comment about content by approximately 25% across all five state-funded channels. When we looked at the relationship for each news outlet, we found that a lower likelihood of receiving a critical comment was statistically significant for AJE (approximately a 35% decrease in likelihood) and RT (approximately a 30% decrease in likelihood). This relationship did not hold for all the channels. While CGTN and VOA News showed no statistical significance, TRT World was associated with a statistically significant higher likelihood of getting a critical comment after the introduction of the YouTube flag—approximately a 100% increase in likelihood.

Additionally, we found that many of the comments in our dataset specifically criticized the channel itself. After the introduction of the label, many critical comments about the source mentioned the policy and discussed the label in the context of government propaganda. We measured and compared the proportion of these critical source comments before and after the implementation of the policy and found that RT was associated with a 70% decrease in likelihood of observing a critical comment about RT as a media source following the introduction of the label (see Table 6).

Next, we controlled for other variables that might have impacted the likelihood of receiving a critical comment. Highly politicized videos, such as those on war and conflict, immigration policy, or the Trump administration, might generate more critical

Table 5. Logit Regression: What is the Likelihood of Getting A Critical Comment (General) After the Policy Implementation?

Critical Comments (General)	Full sample	AJE	CGTN	RT	TRT World	VOA News
Intercept, coefficient (standard error)	-1.01* (0.05)	-0.63* (0.10)	-1.06* (0.27)	-1.41* (0.08)	-0.95* (0.24)	-0.65* (0.10)
After, coefficient (standard error)	-0.28* (0.06)	-0.44* (0.12)	0.01 (0.32)	-0.25* (0.08)	0.74* (0.26)	0.08 (0.14)
After, odds ratio approximate (%)	0.75 (-25%)	0.64 (-35%)	1.12 (+10%)	0.78 (-30%)	2.1 (+100%)	0.02 (+5%)
Sample size (n)	8,071	1,584	230	4,853	438	962

Note. AJE= Al Jazeera English; CGTN=China Global Television Network; RT=Russia Today; VOA News=Voice of America News.

*Indicates the coefficient estimate is statistically significant at the 99% confidence level 3.

comments compared to topics like sports, lifestyle news, or entertainment. Thus, we performed a third logit regression (see Table 7) asking: does the YouTube policy change have an impact on the likelihood that a comment will be critical after controlling for how political a video is? This analysis allowed us to control for the possibility that political videos were not randomly distributed before and after the event. In the first regression, YouTube’s policy change was associated with approximately a 25% lower likelihood of a critical comment. After controlling for how political the video was, the likelihood of receiving a critical comment after the implementation of YouTube’s policy remained unchanged. However, AJE and VOA News were associated with a statistically significant likelihood of getting a critical comment if the video was classified as highly political. For AJE, there was approximately a 20% decrease in likelihood of observing a critical comment if the video was highly political. For VOA News there was an approximately 70% decrease in the likelihood of observing a critical comment. These findings will be discussed in greater detail in the next section of the article.

Discussion

How did YouTube’s state-funding labels affect user engagement with content from these channels? And how did YouTube’s state-funding label affect user comments about these channels and the content they produce? Results varied across our channels, which could be explained by the editorialization of these channels that attract different segments of YouTube users. First, the labels had no impact on the amount of engagement that the videos received before and after the policy introduction, except for RT, where there was a statistically significant decrease in the average of likes that the videos received after the policy introduction. This is worth noting in relation to the analysis on comments by users, where we found a 30% decrease in the likelihood of

Table 6. Logit Regression: What is the Likelihood of Getting A Critical Comment (Source) After the Policy Implementation?

Critical Comments (Source)	Full sample	AJE	CGTN	RT	TRT World	VOA News
Intercept, coefficient (standard error)	-2.60* (0.08)	-2.82* (0.20)	-2.37* (0.43)	-2.27* (0.11)	-2.60* (0.42)	-3.71* (0.31)
After, coefficient (standard error)	-0.46* (0.10)	-0.14 (0.24)	-0.25 (0.52)	-1.14* (0.14)	1.10* (0.44)	0.15 (0.41)
After, odds ratio approximate (%)	0.63 (-35%)	0.87 (-10%)	0.78 (-20%)	0.31* (-70%)	3.0 (+200%)	1.16 (+15%)
Sample size (n)	8,071	1,584	230	4,853	438	962

Note. AJE= Al Jazeera English; CGTN=China Global Television Network; RT=Russia Today; VOA=Voice of America.
*Indicates the coefficient estimate is statistically significant at the 99% confidence level 3.

observing a critical comment after the policy change and a 70% decrease in the likelihood of observing a critical comment about the source specifically. On the one hand, RT received fewer likes after the policy change, but on the other hand, users also became less critical of the content in their comments. While it seems like these two findings do not comport, we suggest that they actually capture two different effects of the policy change on two different audience segments of RT.

Research on RT has demonstrated that the channel targets audiences who share anti-Western and antiestablishment ideologies (Elsawah & Howard, 2020). At the same time, RT is the most popular state-funded news channel, with the most YouTube engagement, especially compared to the other four channels in the study. RT also spends a significant amount of funds in growing audiences and disseminating news content on social media specifically targeted toward American audiences (Bradshaw et al., forthcoming). Thus, people who watch RT might not only belong to an ideologically similar subgroup of users but might stumble upon its content through YouTube’s recommendation algorithm or click-throughs from other social media websites like Twitter or Facebook. Users who do not belong to RT’s dedicated audience base might be less invested in the content and less likely to comment or engage in discussion about its coverage. However, they might engage in smaller actions, such as liking (or disliking) the video. Thus, the flag might have discouraged broader viewers from liking RT content when the state-funding label was presented. For RT’s dedicated audience base, who are more likely to comment on videos, the introduction of YouTube’s state-funding flag could have challenged their worldview. To avoid the cognitive dissonance of the label, belief perseverance could explain the drop in critical comments about the channel and its content following the introduction of the labels.

Table 7. Linear Regression: What is the Likelihood of Getting a Critical Comment When Controlling for the Politicization of Video Content?

Critical Comments with Politicization Control	Full sample	AJE	CGTN	RT	TRT World	VOA News
Intercept, coefficient (standard error)	-0.96* (0.12)	-1.03* (0.16)	-1.03* (0.34)	-11.33 (196.97)	-2.87* (1.08)	0.48 (0.44)
After, coefficient (standard error)	-0.32* (0.06)	-0.48* (0.12)	-0.11 (0.33)	-0.23* (0.09)	0.72* (0.27)	-0.04 (0.14)
After, odds ratio approximate (%)	0.72 (-25%)	0.62 (-40%)	1.11 (+11%)	0.79 (-20%)	2.06 (+105%)	0.90 (-10%)
Highly political, coefficient (standard error)	0.03* (0.07)	-0.79* (0.16)	-0.33 (0.40)	9.90 (196.97)	2.06 (1.06)	-1.14* (0.43)
Highly political, odds ratio	0.75	0.81	0.71	2,001.88	7.84	0.32
Sample size (n)	8,071	1,584	203	4,853	438	962

Note. AJE = Al Jazeera English; CGTN = China Global Television Network; RT = Russia Today; VOA News = Voice of America News.

*Indicates the coefficient estimate is statistically significant at the 95% confidence level 3.

In contrast to RT, TRT World was an outlier in the dataset, where the likelihood of seeing a critical comment increased by 100% following the introduction of YouTube’s policy change. Unlike RT, TRT World is not as widely recognized as being under the editorial control of the Turkish government (Stalinsky, 2019). Because TRT World is also, by comparison, a relatively new channel, its audience base might not be as established as that of a channel like RT which has spent years tailoring content to a particular kind of audience. Thus, the introduction of YouTube’s policy might have made the average YouTube user more critical of the content coming from TRT World.

Finally, across the entire dataset, the analysis showed there was a 25% decrease in the likelihood of observing a critical comment after the policy change. After controlling for how political the videos were, this relationship did not hold for AJE and VOA News, where both channels were associated with a statistically significant decrease in critical comments if the video was classified as highly political. This is a notable finding for these two channels that have greater levels of independence and adherence to professional standards of journalism, because it suggests that users do not necessarily care about the label, but instead care about the kinds of content and coverage these channels provide. Thus, concerns that the state-funded labels would decrease users’ trust in public broadcasters or other state-funded channels that have greater levels of editorial independence are not supported by our findings.

Study Limitations and Future Research

There are several important limitations to this research that should moderate the findings. First, it is important to note that YouTube's state-funded labeling policy is not global. The policy was first introduced in the United States on February 2, 2018 but has subsequently been rolled out in 21 other countries. Given that there is no way to firmly connect a YouTube user to a geographic location, it is not possible to know that all of the comments in the dataset came from people who saw the flag. While channels were selected because of their focus on U.S. politics and global reach, not all of the comments in the dataset would have come from individuals who saw the flag. Notices could also have a different impact on different audiences, and further research should be done to evaluate the impact of state-funded flags in non-Western contexts.

Second, when it came to sampling channels, we selected prominent channels that could be verified as having the flag accurately applied. However, this has meant that other smaller channels or subchannels were left out of the study. For smaller or more obscure subchannels funded by state-organizations, users might be less aware of how the content they consume is being funded, and the policy might have had a different effect on the audiences of these channels.

Third, there are many factors that lead individuals to agree or disagree with news and information online. Comments and engagement metrics can provide some insight into how people consume political information, but unlike conducting an experiment, we do not have as much information about the audience makeup including demographics or preexisting political beliefs. By examining user-generated content, we can observe genuine user interactions on social media platforms. Here, we have found different effects of labeling across the audiences of state-backed media outlets. The findings of our research can help inform the direction of future experimentation and research, where studies examining audience demographics and political values can enrich our understanding and provide nuance to the efficacy of information labels in the context of state-backed media.

Conclusion

While labeling information seems like a sensible strategy to help users navigate an increasingly complex digital information ecosystem to inoculate users against attempts of persuasive attacks, these policies do not always work as intended. Rather, they might challenge the cognitive comfort of many online users who have selectively chosen exposure to state-backed media channels because of their editorialization. Research on the impact of labeling information on social media has shown both mixed and unintended effects: where information labels have shown a modest reduction in user's sharing and believing false information (Mena, 2020; Pennycook et al., 2018), other studies have found unlabeled information, even when false, is more trusted by users (Pennycook et al., 2020). Also, seeing labels reduces people's overall trust in factually correct and professional journalism (Clayton et al., 2019). While it is beyond the scope of this study to experimentally test the effect of these labels on users' cognition, the comments on RT have shown many examples of belief perseverance among some

users who either trivialized the labels or bolstered support for the channel's content. Given the growing efforts by social media platforms to help users navigate an increasingly complex digital information ecosystem, understanding the effect of these labels is critical in improving platform initiatives to combat the spread of propaganda and disinformation online.

This study contributes to the growing body of empirical work examining the effects of labeling information on users' perceptions and behaviors. By analyzing YouTube comments across five state-funded channels (AJE, CGTN, RT, TRT World, and VOA News) before and after the policy change, we identified some mixed results and unintended consequence. First, different audiences reacted differently to the state-funded label. For RT, while there was a decrease in the number of likes that the videos received after the implementation of YouTube's policy, comments on RT became simultaneously less critical. For RT's dedicated audience base—a specific subgroup of users who already do not trust mainstream media—there appeared to be belief perseverance among them. These users reduced the dissonant cognitions caused by the labels by devaluating and decreasing its importance. For TRT World, which is not as readily recognized as a propaganda outlet, the introduction of the label increased the likelihood of observing a critical comment. Thus, while YouTube's policy might have contributed to some users being more critical of content from known propaganda outlets, it might have simultaneously led to trivialization by specific audience subgroups.

Second, news outlets, politicians, and activists criticized YouTube's state-funded policy for treating independent public broadcasters the same as known propaganda outlets like CGTN, RT, and TRT World. However, for AJE and VOA News, broadcasters who have a greater level of independence, the likelihood of receiving a critical comment after YouTube's policy was associated with how political the video was, rather than the policy change. This finding aligns with other experimental research that found that providing information about publishers did not affect user's attitudes or beliefs about news (Dias et al., 2020). Similarly, YouTube's policy treats different audiences in the same way. As shown by our findings, this can have mixed consequences. For audiences whose cognitions or viewpoints are challenged, labels might trigger cognitive dissonance and strategies for its reduction.

In conclusion, YouTube's state-funding labeling practices have had mixed results. Despite various studies highlighting unintended consequences of labeling practices, platforms continue to implement these policies. This is problematic for two reasons. Firstly, one reason labeling is an attractive solution is because it puts the onus on the user to decide what information is good or bad and absolves social media platforms of editorial responsibility. Given their power to shape and influence how people consume and internalize news and information, platforms must be more transparent and ensure greater levels of oversight for their actions and decisions. Secondly, to actually address the spread of propaganda and disinformation online, private self-regulation strategies should be tested and empirically supported. Of course, scientific inquiry takes time and the challenges posed by platforms for our news and information ecosystem need to be imminently addressed. But this limitation could be overcome if the platforms make better efforts to work with scholars and publish the results of any internal testing they have done.

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

Antonella Perini  <https://orcid.org/0000-0002-6526-1784>

Supplemental Material

Supplemental material for this article is available online.

Note

1. We verified that all five state-funded outlets received a label by reading through tech news coverage about the policy change and exchanged emails with an U.S. journalist investigating the accuracy of YouTube's policy implementation. Correspondence with a ProPublica reporter who provided a list of state-funded channels that did not have the YouTube label added.

References

- Aldous, K. K., An, J., & Jansen, B. J. (2019). *View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations*. Proceedings of the International AAAI Conference on Web and Social Media (Vol. 13, pp. 47–57). <https://ojs.aaai.org/index.php/ICWSM/article/view/3208>
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236.
- Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, 39(6), 1037–49.
- Anglin, S. M. (2019). Do beliefs yield to evidence? Examining belief perseverance vs. change in response to congruent empirical findings. *Journal of Experimental Social Psychology*, 82, 176–99.
- Bai, J., Kong, Q., Li, L., Wang, L., & Zeng, D. (2019). Exploring cognitive dissonance on social media. *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 143–145).
- Berelson, B., & Steiner, G. A. (1964). *Human behavior: An inventory of scientific findings*. Harcourt, Brace & World.
- Bradshaw, S., & Howard, P. N. (2018). *Why does junk news spread so quickly across social media?* Knight Foundation. https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/142/original/Topos_KF_White-Paper_Howard_V1_ado.pdf.
- Bradshaw, S., DiResta, R., & Miller, M. (2022). Playing Both Sides: Russian State-Backed Media Coverage of the #BlackLivesMatter Movement. *International Journal of Press & Politics*. <https://doi.org/10.1177/19401612221082052>

- Bullock, J. G., Gerber, A. S., Hill, S. J., & Huber, G. A. (2013). *Partisan bias in factual beliefs about politics*. National Bureau of Economic Research. Retrieved January 24, 2021, from <https://www.nber.org/papers/w19080>
- Chaffee, S. H., & Miyo, Y. (1983). Selective exposure and reinforcement hypothesis: An inter-generational panel study of the 1980 presidential campaign. *Communication Research*, 10(1), 3–36.
- Charmaz, K. (2006). *Constructing grounded theory*. Sage Publications.
- Cinelli, M., Brugnoli, E., Schmidt, A. L., Zollo, F., Quattrocioni, W., & Scala, A. (2020). Selective exposure shapes the facebook news diet. *PLoS One*, 115(3), e0229129.
- Clayton, K., Blair, S., Busam, J.A., & Forstner, S. (2019). John Glance, Guy Green, Anna Kawata 2019. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4), 1073–1095. <https://doi.org/10.1007/s11109-019-09533-0>
- Cobb, M. D., Nyhan, B., & Reifler, J. (2013). Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychology*, 34(3), 307–326.
- Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin*, 26(9), 1151–1164.
- Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One*, 12(5), e0175799.
- Cotton, J. L. (1985). Cognitive dissonance in selective exposure. In J. Bryant & D. Zillmann (Eds.), *Selective exposure to communication* (pp. 11–13). Routledge.
- Cotton, J. L., & Hieser, R. A. (1980). Selective exposure to information and cognitive dissonance. *Journal of Research in Personality*, 14(4), 518–527.
- Davis, J. L., & Jurgenson, N. (2014). Context collapse: Theorizing context collisions and collisions. *Information, Communication & Society*, 17(4), 476–85.
- Diamond, L., Plattner, M. F., & Walker, C. (2016). *Authoritarianism goes global: The challenge to democracy*. JHU Press.
- Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, 1(1). <https://misinformreview.hks.harvard.edu/article/emphasizing-publishers-does-not-reduce-misinformation/>
- Duncan, M. (2020). What's in a label? Negative credibility labels in partisan news. *Journalism & Mass Communication Quarterly*, 99(2), 390–413. <https://doi.org/10.1177/1077699020961856>
- Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, 18(3), 570–578.
- Elsawah, M., & Howard, P. N. (2020). “Anything that causes chaos”: The organizational behavior of Russia Today (RT). *Journal of Communication*, 70(5), 623–45.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.
- Freelon, D., & Wells, C. (2020). Disinformation as political communication. *Political Communication*, 37(2), 145–56.
- Gao, M., Xiao, Z., Karahalios, K., & Fu, W.-T. (2018). To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. Proceedings of the ACM on Human-Computer Interaction 2 (CSCW) (Vol.55, pp. 1–16). <https://dl.acm.org/doi/10.1145/3274324>

- Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users¹. *Journal of Computer-Mediated Communication*, 14(2), 265–285.
- Gleicher, N. (2020). Labeling state-controlled media on Facebook. *About Facebook*. Retrieved July 1, 2020, from <https://about.fb.com/news/2020/06/labeling-state-controlled-media/>
- Ha, A. (2020). Twitter adds labels for government officials and state-controlled media. *Tech Crunch*. Retrieved October 3, 2020, from <https://techcrunch.com/2020/08/06/twitter-state-media-labels/>
- Ha, L., Andreu Perez, L., & Ray, R. (2021). Mapping recent development in scholarship on fake news and misinformation, 2008 to 2017: Disciplinary contribution, topics, and impact. *American Behavioral Scientist*, 65(2), 290–315.
- Hardyck, J. A., & Kardush, M. (1968). A modest modish model for dissonance reduction. In R. Abelson, E. Aronson, W. McGuire, T. Newcomb, M. Rosenberg, & P. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook* (pp. 684–692). Rand McNally.
- Hart, W., Albarracin, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4), 555–588.
- Jeong, M., Zo, H., Lee, C. H., & Ceran, Y. (2019). Feeling displeasure from online social media postings: A study using Cognitive Dissonance Theory. *Computers in Human Behavior*, 97, 231–240.
- Johnson, T. J., & Kaye, B. K. (2013). The dark side of the boon? Credibility, selective exposure and the proliferation of online sources of political information. *Computers in Human Behavior*, 29(4), 1862–1871.
- Jost, J. T. (2015). Resistance to change: A social psychological perspective. *Social Research*, 82(3), 607–636.
- Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, 66, 236–247.
- Kim, A., & Dennis, A. R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, 43(3), 1025–1039.
- Kim, C., & Yang, S.-U. (2017). Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review*, 43(2), 441–449.
- Kofman, A. (2019). YouTube promised to label state-sponsored videos but doesn't always do so. *ProPublica*. Retrieved January 24, 2021, from <https://www.propublica.org/article/youtube-promised-to-label-state-sponsored-videos-but-doesnt-always-do-so?token=KV1iffz1F9fK0zFiLSbSXrhlEeeah5Bq>
- Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., & Schudson, M. (2018). The science of fake news. *Science*, 359(6380), 1094–1096.
- Lim, J. S., & Ki, E.-J. (2007). Resistance to ethically suspicious parody video on YouTube: A test of inoculation theory. *Journalism & Mass Communication Quarterly*, 84(4), 713–28.
- Lyons, T. (2017). Replacing disputed flags with related articles. *Facebook Newsroom*. <https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/>.
- Masip, P., Suau-Martínez, J., & Ruiz-Caballero, C. (2018). Questioning the selective exposure to news: Understanding the impact of social networks on political news consumption. *American Behavioral Scientist*, 62(3), 300–319.
- McFaul, M., & Kass, B. (2019). Understanding Putin's intentions and actions in the 2016 US presidential election. In M. McFaul (Ed.), *Securing American elections: Prescriptions for enhancing the integrity and independence of the 2020 U.S. presidential election and beyond* (pp. 108). Stanford Cyber Policy Center.

- Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & Internet*, 12(2), 165–83.
- Mohan, N., & Kyncl, R. (2018). Building a better news experience on YouTube, together. *Official YouTube Blog*. Retrieved July 3, 2020, from <https://youtube.googleblog.com/2018/07/building-better-news-experience-on.html>
- Molter, V. (2020). Virality project (China): Pandemics and propaganda. Retrieved February 1, 2021, from <https://cyber.fsi.stanford.edu/news/chinese-state-media-shapes-coronavirus-convo>
- Mosseri, A. (2017). A new educational tool against misinformation. *Facebook Newsroom*. <https://newsroom.fb.com/news/2017/04/a-new-educational-tool-against-misinformation/>
- Muntinga, D. G., Moorman, M., & Smit, E. G. (2011). Introducing COBRAs: Exploring motivations for brand-related social media use. *International Journal of Advertising*, 30(1), 13–46.
- Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of Foreign misinformation. *Harvard Kennedy School Misinformation Review*. Retrieved January 24, 2021, from <https://misinforeview.hks.harvard.edu/article/state-media-warning-labels-can-counteract-the-effects-of-foreign-misinformation/>
- Nestler, S. (2010). Belief perseverance: The role of accessible content and accessibility experiences. *Social Psychology*, 41(1), 35–41.
- Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2019). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, 42, 939–960.
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303–30.
- Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11), 4944–4957.
- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865–80.
- Pfau, M., Tusing, K. J., Koerner, A. F., Lee, W., Godbold, L. C., Penaloza, L. J., Yang, V. S. H., & Hong, Y. H. (1997). Enriching the inoculation construct: The role of critical components in the process of resistance. *Human Communication Research*, 24(2), 187–215.
- Quattrociochi, W., Scala, A., & Sunstein, C. R. (2016). *Echo chambers on Facebook*. Social Science Research Network. Retrieved February 9, 2021, from <https://papers.ssrn.com/abstract=2795110>
- Radsch, C. (2018). YouTube labels on public broadcasters draw Ire in US, Russia. *Committee to Protect Journalists*. Retrieved January 24, 2021, from <https://cpj.org/2018/03/youtube-labels-on-public-broadcasters-draw-ire-in/>
- Roth, Y., & Achuthan, A. (2020). *Building rules in public: Our approach to synthetic & manipulated media*. Retrieved July 3, 2020, from https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html
- Roth, Y., & Pickles, N. (2020). *Updating our approach to misleading information*. Retrieved July 3, 2020, from https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html
- Samantha, B., DiResta, R., & Miller, M. (2022). Playing Both Sides: Russian State-Backed Media Coverage of the #BlackLivesMatter Movement. *International Journal of Press & Politics*. <https://doi.org/10.1177/19401612221082052>
- Samek, G. (2018). Greater transparency for users around news broadcasters. *Official YouTube Blog*. Retrieved June 19, 2020, from <https://youtube.googleblog.com/2018/02/greater-transparency-for-users-around.html>

- Schmid, P., & Betsch, C. (2019). Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour*, 3(9), 931–39.
- Schreiner, M., Fischer, T., & Riedl, R. (2019). Impact of content characteristics and emotion on behavioral engagement in social media: Literature review and research agenda. *Electronic Commerce Research*, 21, 329–345. <https://doi.org/10.1007/s10660-019-09353-8>
- Shaban, H. (2018). How public media could become a casualty of YouTube's war on propaganda. *Washington Post*. <https://www.washingtonpost.com/news/the-switch/wp/2018/02/09/how-public-media-could-become-a-casualty-of-youtubes-war-on-propaganda/>
- Skoric, M. M., Zhu, Q., & Tammy Lin, J.-H. (2018). What predicts selective avoidance on social media? A study of political unfriending in Hong Kong and Taiwan. *American Behavioral Scientist*, 62(8), 1097–1115.
- Stalinsky, S. (2019). This Turkish TV network is Erdogan's propaganda arm. U.S. experts should stop appearing on it. *Washington Post*. Retrieved June 29, 2020, from <https://www.washingtonpost.com/opinions/2019/11/12/how-us-can-send-message-erdogan-free-press/>
- Swani, K., & Labrecque, L. I. (2020). Like, comment, or share? Self-presentation vs. brand relationships as drivers of social media engagement choices. *Marketing Letters*, 31(2), 279–98.
- Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *PsyArXiv Preprint*. Retrieved July 12, 2020, from <https://osf.io/ba2kc>
- Taylor, E., Walsh, S., & Bradshaw, S. (2018). *Industry responses to the malicious use of social media*. NATO STRATCOM COE.
- Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3), 460–80.
- Twitter Safety. 2019. *Defining public interest on Twitter*. Retrieved July 3, 2020, from https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html
- Vaidis, D. C., & Bran, A. (2018). Some prior considerations about dissonance to understand its reduction: Comment on McGrath (2017). *Social and Personality Psychology Compass*, 12(9). <https://doi.org/10.1111/spc3.12411>
- Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2), 1600008.
- Weeks, B. E., & de Zúñiga, H. G. (2021). What's next? Six observations for the future of political misinformation research. *American Behavioral Scientist*, 65(2), 277–89.
- Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1), 135–63.

Author Biographies

Samantha Bradshaw is an Assistant Professor at American University's School of International Service.

Mona Elswah is a Lecturer at Cairo University.

Antonella Perini is a Research Associate at the Alan Turing Institute.