Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation

Shan Jiang, Ronald E. Robertson, Christo Wilson
Northeastern University, USA
{sjiang, rer, cbw}@ccs.neu.edu,

Abstract

Social media platforms have been the subject of controversy and scrutiny due to the spread of hateful content. To address this problem, the platforms implement content moderation using a mix of human and algorithmic processes. However, content moderation itself has lead to further accusations against the platforms of political bias. In this study, we investigate how channel partisanship and video misinformation affect the likelihood of comment moderation on YouTube. Using a dataset of 84,068 comments on 258 videos, we find that although comments on right-leaning videos are more heavily moderated from a correlational perspective, we find no evidence to support claims of political bias when using a causal model that controls for common confounders (e.g., hate speech). Additionally, we find that comments are more likely to be moderated if the video channel is ideologically extreme, if the video content is false, and if the comments were posted after a fact-check.

1 Introduction

In the wake of 2016–2017 global election cycle, social media platforms have been subject to heightened levels of controversy and scrutiny. Besides well-documented privacy issues (Solon 2018), social media platforms have been increasingly used to promote partisanship (Allcott and Gentzkow 2017), spread misinformation (Constine 2018), and breed violent hate speech (Olteanu et al. 2018).

The solution promulgated by social media platforms for these ills is an increase in *content moderation*. In terms of mechanisms, the major platforms have committed to hiring tens of thousands of new human moderators (Levin 2017), investing in more artificial intelligence to filter content (Gibbs 2017), and partnering with fact-checking organizations to identify misinformation (Glaser 2018). In terms of policy, the platforms are updating their *community guidelines* with expanded definitions of hate speech, harassment, etc (YouTube 2018; Facebook 2018; Twitter 2018).

This increased reliance on content moderation faces a backlash from ideological conservatives, who claim that social media platforms are biased against them and are censoring their views (Kamisar 2018; Usher 2018). Two US House Committees have held hearings on content moderation practices to "specifically look at concerns regarding a lack of transparency and potential bias in the filtering practices of social media companies (Facebook, Twitter and YouTube)" (Bickert, Downs, and Pickles 2018; Dorsey 2018). These concerns are driven by multiple factors, including anecdotal reports that: Facebook's Trending News team did not promote stories from conservative media outlets (Nunez 2016), Twitter "shadow banned" conservative users (Newton 2018), fact-checking organizations are biased (Richardson 2018), and selective reporting by partisan news agencies (Arceneaux, Johnson, and Murphy 2012).

However, there is, to our knowledge, no scientific evidence that social media platforms' content moderation practices exhibit systematic partisan bias. On the contrary, there are many cases where ideologically liberal users were moderated, although these cases have received less attention in the media (Masnick 2018). It is possible that moderation only appears to be biased because political valence is correlated with other factors that trigger moderation, such as bullying, calls to violence, or hate speech (Gillespie 2018). Further, there is evidence suggesting that users tend to overestimate bias in moderation decisions (Shen et al. 2018).

In this study, we use YouTube as a lens and take a first step towards disentangling these issues by investigating how partisanship and misinformation in videos affect the likelihood of comment moderation. Specifically, we examine four hypotheses related to four variables of YouTube videos: the direction of partisanship (**H1a$_0$: left/right**), the magnitude of partisanship (**H1b$_0$: extreme/center**), the veracity of the content (**H2a$_0$: true/false**), and whether a comment was posted before or after the video was fact-checked (**H2b$_0$: before/after**). For each variable, we start with the null hypotheses (**H$_0$**) that the variable has no effect on comment moderation, and then use correlational and casual models to collect evidence on rejecting the null hypotheses. The conceptual framework of our hypotheses is shown in Figure 1.

To investigate these hypotheses, we collected a dataset of 84,068 comments posted across 258 YouTube videos,[1] and associate them to partisanship labels from previous

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]The dataset is available at: https://moderation.shanjiang.me

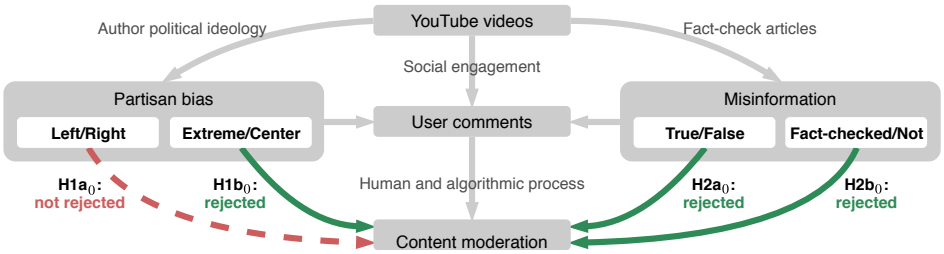

Figure 1: **Conceptual framework.** We investigate the effect of partisanship (i.e., left/right, extreme/center) and misinformation (i.e., true/false, fact-checked/not) on comment moderation. Potential confounders include social engagement on YouTube videos (e.g., views and likes) and linguistics in comments (e.g., hate speech).

research (Robertson et al. 2018) and misinformation labels from Snopes or PolitiFact (Jiang and Wilson 2018). We first run a correlational analysis and find that all of our hypothesized variables significantly correlate with the likelihood of comment moderation, suggesting a political bias against right-leaning content. However, we argue that such bias is *misperceived* as it ignores other confounding variables that potentially contribute to moderation decisions, such as social engagement (e.g., views and likes) (Mondal, Silva, and Benevenuto 2017) and the linguistics in comments (e.g., hate speech) (Shen et al. 2018; Chandrasekharan et al. 2018). Therefore, we re-analyze our dataset using a causal propensity score model to investigate null hypotheses when potential confounds are controlled. We make the following observations:

- **H1a$_0$: not rejected.** No significant difference is found for comment moderation on left- and right-leaning videos.
- **H1b$_0$: rejected.** Comments on videos from ideologically extreme channels are $\sim$50% more likely to be moderated than center channels.
- **H2a$_0$: rejected.** Comments on true videos are $\sim$60% less likely to be moderated than those on false videos.
- **H2b$_0$: rejected.** Comments posted after a video is fact-checked are $\sim$20% more likely to be moderated than those posted before the fact-check.

We approach these hypotheses using an empirical method for auditing black-box decision-making processes (Sandvig et al. 2014) based on publicly available data on YouTube. Neither we, nor the critics, have access to YouTube's internal systems, data, or deliberations that underpin moderation decisions. Instead, we aim to highlight the difference in *perceived* bias when analyzing available data using correlational and causal models, and further, foster a healthier discussion of algorithmic and human bias in social media.

2 Background & Related Work

Social media platforms publish sets of *community guidelines* that explain the types of content they prohibit in order to guide their content moderation practice (YouTube 2018; Facebook 2018; Twitter 2018). In the case of YouTube, it lists rules for: nudity or sexual content, harmful or dangerous content, hateful content, violent or graphic content, harassment and cyberbullying, etc (YouTube 2018). Once content on YouTube (e.g., a video or comment) is judged to violate the guidelines, it is taken down, i.e., *moderated*.

There are multiple reasons why a comment could be moderated on YouTube. A comment may be reviewed by patrolling YouTube moderators, or a comment may be *flagged* by YouTube users and then reviewed by the YouTube moderators (Levin 2017). Additionally, a comment may be removed by the corresponding video uploader, or by the commenter themselves (YouTube 2018). Besides these human efforts, YouTube also uses algorithms that automatically flag and moderate inappropriate content (Gibbs 2017). In general, the mechanisms that lead to comment moderation are convoluted. Therefore, we view the internal YouTube system as a black-box, and focus on the *outcome* of moderation instead.

Pros & Cons of Content Moderation

Content moderation has been shown to have positive effects on social media platforms. A study that investigated Reddit's ban of the r/fatpeoplehate and r/CoonTown communities found that the ban expelled more "bad actors" than expected, and those who stayed posted much less hate speech than before the ban (Chandrasekharan et al. 2017). A study that interviewed users of Twitter's "blocklist" feature discussed how it can be used to prevent harassment (Jhaver et al. 2018).

However, content moderation systems have also raised concerns about bias and efficacy. Human moderators have been shown to bring their own biases into the content evaluation process (Diakopoulos and Naaman 2011) and automated moderation algorithms are prone to false positives and negatives (Veletsianos et al. 2018). These moderation strategies are also brittle: a study on Instagram found that users in pro-eating disorder communities invented variations of banned tags (e.g., "anorexie" instead of "anorexia") to circumvent lexicon-based moderation (Chancellor et al. 2016).

Researchers have also studied the community norms behind moderation from a linguistic perspective. A study on Reddit used 2.8M removed comments to identify macro-, meso-, and micro-norms across communities (Chandrasekharan et al. 2018). A study on the Big Issues Debate group of Ravelry found that comments expressing unpopular viewpoints were more likely to be moderated, but that this effect is negligible when compared to the total level of moderation (Shen et al. 2018). These studies highlight the role of linguistics on the task of comment moderation, which sheds

light on the importance of controlling for linguistics when investigating bias in moderation practices.

Algorithmic & Human Bias on the Web

The opaque nature of online systems has led to investigation on whether they exhibit bias against specific groups (Sandvig et al. 2014). Studies have found gender and racial bias on hiring sites (Chen et al. 2018), freelance markets (Hannák et al. 2017), ridesharing platforms (Jiang et al. 2018), and online writing communities (Fast, Vachovsky, and Bernstein 2016). In the case of ideological groups, it has been reported that social media platforms such as Facebook are inferring users' ideologies to target them with political ads (Speicher et al. 2018), while search engines may create "filter bubbles" that isolate users from ideologically opposing information (Liao and Fu 2013; Hu et al. 2019).

However, research on ideological bias in online contexts has sometimes led to surprising conclusions. Facebook researchers found that the partisan bias of content appearing in the Newsfeed was due more to homophily than algorithmic curation (Bakshy, Messing, and Adamic 2015). A study on Google Search also found that the partisan bias of search results was dependent largely on the input query rather than the self-reported ideology of the user (Robertson et al. 2018).

As for content moderation, there have been several claims that social media platforms are censoring or biased against political conservatives (Kamisar 2018; Usher 2018). In August 2018, the 45th President of the United States stated that tech companies "are totally discriminating against Republican/Conservative voices", though no evidence was offered to back the claim (Murray and Lima 2018).

In this study, $H1a_0$ and $H1b_0$ aim to study if, and to what extent, claims of political bias against YouTube's comment moderation system are justified by investigating the role of partisanship on comment moderation. Furthermore, because online misinformation campaigns have been to shown to have high correlation with partisanship (Allcott and Gentzkow 2017), we also investigate the $H2a_0$ and $H2b_0$, i.e., role of misinformation on comment moderation in YouTube.

3 Data Collection

To answer our hypotheses, we leverage a dataset of 84,068 comments posted on 258 YouTube videos, along with labels including *outcome* (was a comment moderated), *treatments* (corresponding to our four hypothesized variables), and *controls* for confounding variables (i.e., social engagement and the linguistic features of comments). In this section, we describe our data collection and labeling methods with an illustrative example in Figure 2. Summary statistics are shown in Table 1.

Dataset & Moderation Outcome

We start with an initial dataset collected from previous research for analyzing user comments under misinformation (Jiang and Wilson 2018). Jiang and Wilson crawled Snopes and PolitiFact in January 2018, identified all fact-check articles that linked to posts on social media, including videos on YouTube, and then crawled all the comments attached to these posts. This dataset contains over 2K YouTube videos with 828K comments and is, to the best of our knowledge, the only available dataset of YouTube comments with veracity labels for videos. Figure 2 shows an example article from PolitiFact (Snyder 2016) that fact-checked a YouTube video from Red State Media (RedStateMedia 2016).

Table 1: **Statistics of dataset.** Mean with 95% confidence intervals after labeling are shown for each measured variable.

Type	Variable	Value	Mean ± 95% CI
Outcome	Moderated/Not	1/0	0.032 ±0.001
Misinformation	True/False	1/0	0.132 ±0.002
	After/Before Fact-check		0.332 ±0.003
Partisan bias	Right/Left	1/0	0.472 ±0.003
	Extreme/Center		0.716 ±0.003
Engagement	Views	0-3	1.407 ±0.008
	Likes		1.438 ±0.007
	Dislikes		1.411 ±0.008
Linguistic	Swear	1/0	0.102 ±0.002
	Laugh		0.052 ±0.002
	Emoji		0.024 ±0.001
	Fake		0.086 ±0.002
	Administration		0.041 ±0.001
	American		0.022 ±0.001
	Nation		0.016 ±0.001
	Personal		0.239 ±0.003

To determine whether each comment in the dataset was moderated (1) or not (0), we recrawled all of the YouTube videos in June 2018. We label comments that appeared in the first crawl but not the second as *moderated*. There are two limitations of this labeling method: a) we do not know why or who moderated each comment, and we discuss this limitation more deeply in later sections; and b) our dataset only contains comments that were moderated after January and before June 2018. Figure 2 shows four example comments from our dataset, two of which were moderated.

Partisanship Treatments

We use two measures for partisanship: its direction (i.e., left (0) or right (1)) for $H1a_0$ and magnitude (i.e., extreme (1) or center (0)) for $H1b_0$ of each video in our dataset. This information is not contained in the original dataset (Jiang and Wilson 2018). To gather this information, we leverage partisan scores from previous research (Robertson et al. 2018). In brief, these scores were constructed using a virtual panel of registered US voters. Voters were linked to their Twitter accounts, and then the partisan score of a website was measured by the relative proportion of how it was shared by Democrats and Republicans. This dataset contains scores for 19K websites and the scores range from -1 (shared entirely by Democrats) to 1 (shared entirely by Republicans).

Since the basic unit of our analysis is YouTube videos, not websites, we used Google Search as an intermediary to link a YouTube channel to its website. We entered all 19K website domains as queries into Google Search and added a filter to only return results from the YouTube domain. For

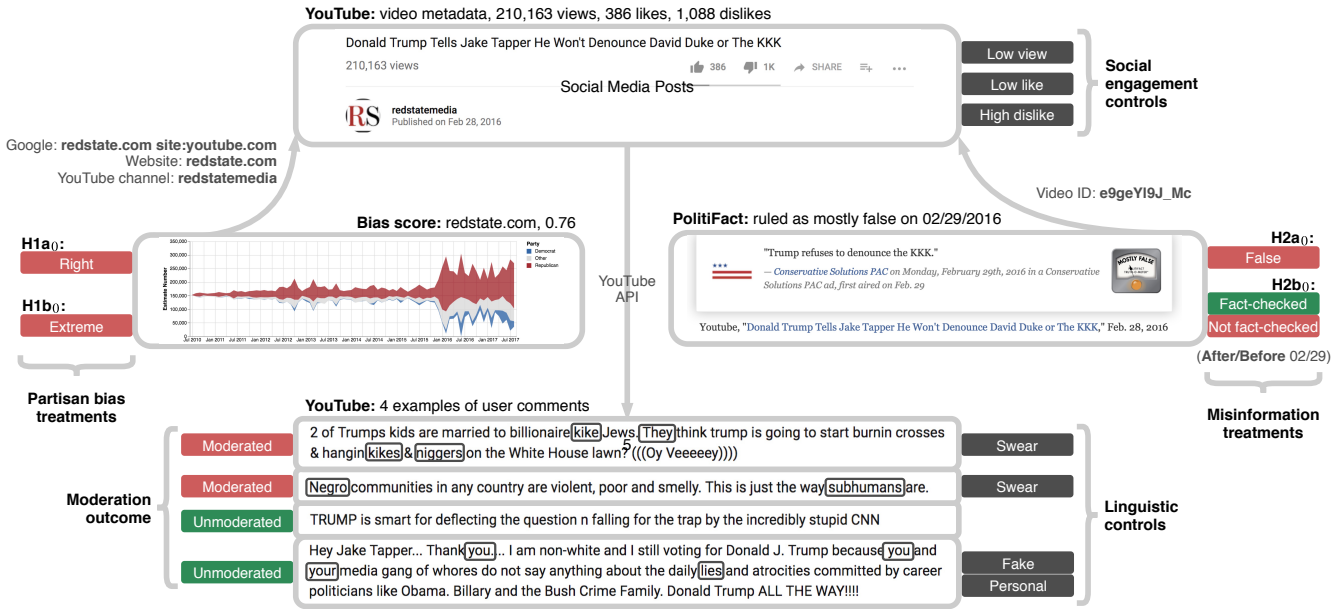

Figure 2: **Data collection process and an illustrative example.** Starting from a fact-check article on PolitiFact, we collect the misinformation treatment and a YouTube video ID. Another starting point is the partisan score for the website "redstate.com", where we collect the partisanship treatment and then use Google to get the corresponding channel name. We then use YouTube API to collect the video metadata and link previous data by video ID and channel name respectively. We also collect user comments and labeled their linguistic treatments using *ComLex*. Finally, we compare two crawls to identify moderated comments.

each query, we located the first search result containing a link to a YouTube channel (if one existed on the first page of search results), and compared the ID of that channel to the IDs of all channels in our dataset. If we found a match, we associated the partisan score of that website to videos in our dataset from that channel.

Using this process, we were able to associate partisanship labels to 258 YouTube videos from our dataset, originating from 91 unique channels. Example channels include "MacIverInstitute", "John McCain", "BarackObamadotcom", etc. The remaining videos were posted by users and channels that had little-to-no presence off of YouTube. For direction of partisanship, we label each video as *left* or *right* depending on whether its partisanship score is < 0 or > 0, respectively. Further, for magnitude of partisanship, we labels each video as *extreme* or *center* depending on whether the absolute value of its partisanship score is > 0.5 or < 0.5, respectively.[2]

For example, as shown in Figure 2, the partisan score for "redstate.com" is 0.76. We use Google to search the query "redstate.com site:youtube.com" and follow the first link that contains a YouTube channel ID, which leads us to the Red State Media YouTube channel (RedStateMedia 2018). This enables us to label all Red State Media videos in our dataset as *right* and *extreme*.

Misinformation Treatments

We use two measures for misinformation: the veracity of each video (i.e., true (1) or false (0)) for $H2a_0$ and whether each comments was posted before (0) or after (1) the video was fact-checked for $H2b_0$. The dataset from Jiang and Wilson already contains articles from Snopes and PolitiFact with veracity rulings and timestamp.

We label a video as *true* if the corresponding fact-check article determined that it was true, otherwise we label the video as *false*.[3] For *before/after* labels, we compare the timestamp of each comment to the timestamp of the corresponding fact-check article. The example in Figure 2 shows that PolitiFact judged this video to be false on February 29, 2016.

Social Engagement Controls

We also collected social engagement information (i.e., views, likes, and dislikes) as potential controls, e.g., a video with many dislikes could attract more flaggers and therefore cause more moderation. We bin the number of views to an integer in the range 0 (low, $< 25\%$ quantile) to 3 (high, $> 75\%$ quantile) based on quantiles of the view distribution. Similarly, we process likes/dislikes by normalizing them with the number of views to get like/dislike rates per video, then bin them in the same manner as views.[4]

[2]We discuss results using alternative thresholds in later sections.

[3]Thus, our binary veracity label encodes the presence or absence of misinformation in a video, regardless of magnitude. We use a binary encoding for veracity because Jiang and Wilson found that users exhibit significantly different linguistic patterns in comments depending on whether misinformation is present.

[4]This step improves the model performance in later sections. Continuous data are vulnerable to outliers, and number of likes/dislikes without normalization shows high multicollinearity with number of views, i.e., highly viewed videos have more likes and

The example video in Figure 2 has 210,163 views, 386 likes (0.184% like rate) and 1,088 dislikes (0.518% dislike rate), which we label as *low view* (25% quantile), *low like* (25% quantile), and *high dislike* (75% quantile).

Linguistic Controls

We use a lexicon-based approach to control for the linguistics of each comment, as linguistics are the primary moderation criteria in YouTube's community guidelines (YouTube 2018) and have been found to affect moderation in practice (Shen et al. 2018; Chandrasekharan et al. 2018).

For this task, we use an existing lexicon called *Comlex* (Jiang and Wilson 2018) that contains 28 categories (56 subcategories) of human evaluated words extracted from user comments on social media, i.e., the same context as our study. Prior work has found that using contextually appropriate lexicons yields better results than generic ones (Li, Lu, and Long 2017).[5] We apply standard text pre-processing techniques to the comments in our dataset using NLTK (Loper and Bird 2002) (e.g., tokenization, case-folding, and lemmatization) before mapping them into ComLex.

We select eight word categories that significantly ($p < 0.001$) affect moderation likelihood for comments, determined by a preliminary linear regression model:[6] *swear* (including hate speech, e.g., "fuck", "bitch", "nigger"), *laugh* (e.g., "lol", "lmao", "hahaha"), *emoji* (e.g., "😂", "😊", "💩"), *fake* (fake awareness, e.g., "lie", "propaganda", "bias"), *administration* (e.g., "mayor", "minister", "attorney"), *American* (cities and states, e.g., "nyc", "texas", "tx"), *nation* (other nations, e.g., "canada", "mexico", "uk"), and *personal* (e.g., "your", "my", "people's").[7] We construct eight binary variables for each comment in our dataset; each variable is 1 if the given comment includes a word from that category.

Figure 2 shows four examples of user comments under the video. The first comment contains the hate lemmas "kike" and "nigger", therefore it is labeled as *swear*. Similarly, the second contains "negro" and "subhuman" so it is also labeled as *swear*. The last comment contains the lemma "lie" which is a word from the *fake* awareness category, and the lemma "your" and "you" which are from the *personal* category, therefore these variables are 1. All other linguistic variables that contains no words are labeled as 0.

Ethics of Data Collection

We obeyed community-standard ethical practices during our data collection. We only collected data from the official YouTube API and respected the service's rate limits (i.e., we did not circumvent them using "sock puppets"). All user IDs have been removed from our public data release.

dislikes. (Original data: Spearman $\rho = 0.949^{***}$ for views/likes, and $\rho = 0.887^{***}$ for views/dislikes. After normalization and binning: $\rho = 0.249^{***}$ for views/likes, and $\rho = -0.625^{***}$ for views/dislikes.)

[5] We also applied generic lexicons such as LIWC. We discuss these results in later sections.

[6] This step is designed to select relevant categories. Including all categories would harm the results of our causal model due to overfitting in the logistic regressions to calculate propensity scores.

[7] The full list of words in the selected categories is available in the Supplementary Materials.

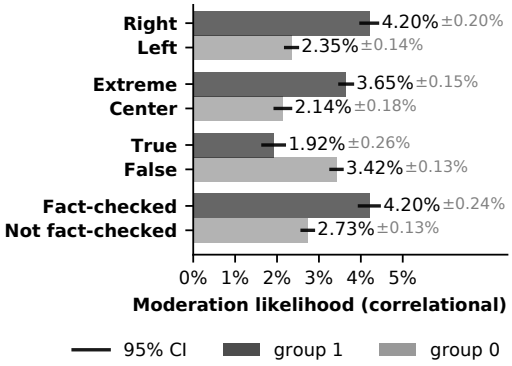

Figure 3: **Correlational difference in moderation likelihood.** Moderation likelihood for each group with 95% CI is shown. All four null hypotheses are rejected.

4 Correlation

In this section, we conduct correlational analysis of our data to investigate the perception of partisan bias in content moderation, and argue that such bias is misperceived.

Correlational Perception of Bias

We frame the correlational perception of bias as the raw difference in moderation likelihood under each hypothesized variable, i.e., if moderation likelihood under one label (e.g., *right*) is significantly different from its dual (*left*), the corresponding null hypothesis is rejected (correlationally) by our dataset. The moderation likelihood under each hypothesized variable with 95% confidence interval (CI) is shown in Figure 3. We perform a χ^2 test on the significance of difference in likelihood between each pair. Under this intuitive, but naïve, perception of bias, all null hypotheses are rejected.

For $H1a_0$, we see that there is a 79% increase in the moderation likelihood on comments from *right*-leaning videos versus *left*-leaning videos, and that the difference is significant ($\chi^2 = 231.0^{***}$).[8] This finding seems to support, at least on the surface level, the claim that content moderation is biased against conservatives (Kamisar 2018; Usher 2018). For $H1b_0$, we observe a 71% increase in moderation likelihood from *center* to *extreme* channels, which is also significant ($\chi^2 = 125.2^{***}$). This observation could be caused by YouTube's efforts to monitor extremely partisan channels to prevent hateful content (News 2017; Morris 2017; Chatterjee and Dave 2017).

For $H2a_0$, we find that there is a 44% decrease in the likelihood that comments will be moderated when moving from *false* to *true* videos, and that this difference is significant ($\chi^2 = 69.6^{***}$). Similarly, for $H2b_0$, we observe a 54% increase in moderation likelihood for comments posted after a fact-check on the associated video is available, which is also significant ($\chi^2 = 129.1^{***}$). These findings may be related

[8] $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

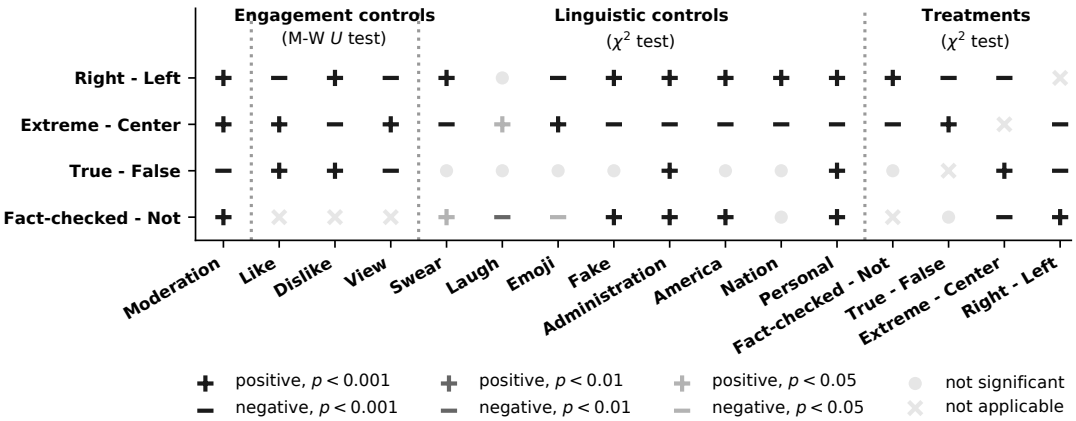

Figure 4: **Correlational difference for confounding variables.** The 1st column repeats the observations we made for moderation likelihood. The 2nd to 4th columns show how social engagement correlates with hypothesized variables, the 5th to 12th columns show linguistic features, and 13th to 16th columns show how hypothesized variables correlate with each other. Each "**+**" represents a positive difference in mean and "**–**" a negative one. Significance, as suggested by χ^2 or Mann-Whitney (M-W) U test, is encoded with transparency.

to YouTube's purported efforts to fight misinformation on their platform (Alvarez 2018; Post 2018; Palladino 2018) by actively partnering with fact-checking organizations (Glaser 2018; McCracken 2018).

Of course, the correlations we report in Figure 3 are potentially specious, since we do not control for correlations between these treatments or with other confounding variables. Therefore, **we do not endorse the findings presented in Figure 3.** Rather, we present these results merely to highlight why a person might erroneously believe that comment moderation on YouTube exhibits partisan bias.

The Problem of Confounding Variables

Comment moderation on YouTube is complicated. As shown in Figure 4, there are a set of potential confounding variables that correlate with our hypothesized variables. The 1st column repeats our observations from Figure 3. The 2nd to 4th columns show how social engagement on videos correlates with the hypothesized variables, while the 5th to 12th columns show correlations with linguistic features. Finally, the 13th to 16th columns examine correlations between the hypothesized variables themselves. Each "**+**" represents a positive difference in mean and "**–**" a negative one. Significance, calculated using the χ^2 or Mann-Whitney (M-W) U test, is encoded with transparency.[9]

Take H1a_0 as an example. With respect to video-level confounders, *right*-leaning videos have significantly less views ($U = 0.310 \cdot 10^{9***}$) and likes ($U = 0.333 \cdot 10^{9***}$), but significantly more dislikes ($U = 0.408 \cdot 10^{9***}$) than *left*-leaning videos. This provides an alternative explanation for the seeming partisan bias of moderation: the higher dislike rate may result in more flagged comments, thus increasing the likelihood of moderation.

[9]Since we present 57 independent χ^2 and M-W U tests, we use Bonferroni correction to counteract the problem of multiple hypothesis testing.

With respect to comment-level linguistics, *right*-leaning videos contain significantly more swear words ($\chi^2 = 671.2^{***}$), fake awareness signals ($\chi^2 = 1013.6^{***}$), discussion on administrative matters ($\chi^2 = 778.5^{***}$), references to city/states in America ($\chi^2 = 686.6^{***}$) and other nations ($\chi^2 = 117.1^{***}$), and personal pronouns ($\chi^2 = 423.7^{***}$), but less usage of emojis ($\chi^2 = 524.9^{***}$). This also provides alternative explanations for the seeming partisan bias of moderation: perhaps comments on *right*-leaning videos are more heavily moderated because they include more hate speech.

We also observe that *right*-leaning videos are significantly more likely to be fact-checked ($\chi^2 = 4738.9^{***}$) and *false* ($\chi^2 = 221.8^{***}$) than *left*-leaning videos. This reveals another complication: our hypothesized variables are correlated with each other. This suggests another alternative explanation for H1a_0: that misinformation is the driving force behind moderation, not partisanship.

Some of the correlations in Figure 4 are supported by findings from existing research. For example, we find no significant difference in fake awareness signals between *true* and *false* videos ($\chi^2 = 8.4, p = 0.004$), which agrees with previous work on people's inability to identify misinformation (Ward et al. 1997; Robinson et al. 1995; Nickerson 1998). Additionally, we observe that comments posted after fact-checking contain more fake awareness signals ($\chi^2 = 149.7^{***}$), which suggests positive effects of fact-checking on people's expression of political beliefs (Porter, Wood, and Kirby 2018; Fridkin, Kenney, and Wintersieck 2015). However, we also observe more swear word usage ($\chi^2 = 12.8^{*}$) which could be linked to "backfire" effects, where attempts to correct false beliefs makes things worse (Nyhan and Reifler 2010; Wood and Porter 2016).

5 Causality

To disentangle the effects of our hypothesized variables, we apply a causal model that controls for identified confounding

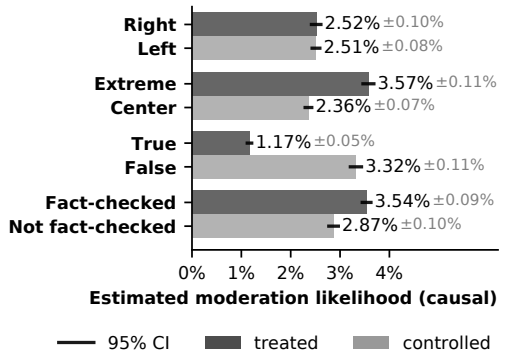

Figure 5: **Causal difference in moderation likelihood.** Moderation likelihood for controlled and treated groups with 95% CI is shown. H1a$_0$ is no longer rejected. Differences in the other 3 hypothesized variables are also changed.

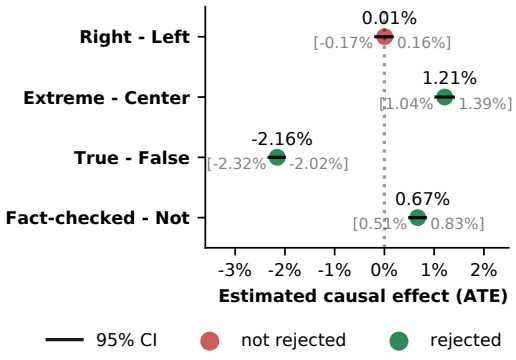

Figure 6: **Causal effect estimation.** Average treatment effect (ATE) with 95% CI is shown. Significance level for null hypothesis is encoded with color. CIs using bootstrap are considered as conservative estimates.

variables. A causal effect is framed as the difference between "what happened" and "what would have happened" (Pearl 2009), e.g., H1a$_0$ is framed as "what would happen if a left-leaning video changed to right-leaning (while its partisanship magnitude, misinformation level, social engagement, etc. remained the same)". One way to estimate causal effects from observational data is called *matching*. The idea is to find *quasi-experiments* where subjects have similar controls but different treatments, and then compare their outcomes.

Several different matching methods have been proposed for causal inference, such as exact matching, Mahalanobis distance, and propensity scoring (Stuart 2010). The latter two have been used within the Computer Science community (Chandrasekharan et al. 2017; Olteanu, Varol, and Kiciman 2017; Chen et al. 2018; Foong et al. 2018). One shortcoming of exact matching and Mahalanobis distance is that the matching is based on each confounding variable, meaning that the number of matches typically decreases as the number of confounders increases. Therefore, we use a propensity scoring method (Rosenbaum and Rubin 1983) that has been used extensively in the social (Thoemmes and Kim 2011), psychological (Lanza, Moore, and Butera 2013), and biological (Austin 2008) literatures.

Model Specification

The propensity score is the *probability of getting the treatment label*. It summarizes all of the confounding variables into one scalar. It has been proven that propensity scores are balancing scores, i.e., given a particular propensity score, the distribution of confounders that yield such a score is the same in the treated and controlled groups. Therefore, matching individuals with similar propensity scores mimics a quasi-experiment, at least for measured confounding variables. Additionally, if such an experiment is randomized given a measured set of confounders, then the treatment assignment is also randomized given the propensity scores, which justifies matching based on the propensity score rather than on the full spectrum of confounders (i.e., exact matching and Mahalanobis distance) (Rosenbaum and Rubin 1983).

For each of our hypotheses, we compute propensity scores using measured confounding variables and the other three hypothesized variables. We then match each treated/controlled sample with its *2-nearest neighbors* based on propensity scores. Finally, we estimate causal effects, denoted as the Average Treatment Effect (ATE), by averaging the difference in mean for each treated/controlled pair and bootstrap CIs and p-values.

Causal Perception of Bias

The estimated mean of each hypothesized variable with 95% CI is shown in Figure 5, where light (dark) bars represent the controlled (treatment) group. The causal effect estimation with 95% bootstrapped CI[10] is shown in Figure 6. We depict H1a$_0$ in red since it is no longer rejected, while we depict the three hypotheses that are still rejected in green.[11]

H1a$_0$ is no longer rejected. In the controlled setting, the estimated moderation likelihood for comments under *left*-leaning videos is $2.51\% \pm 0.08\%$ and under *right*-leaning video is $2.52\% \pm 0.10\%$, which represents an estimated causal effect of 0.01% (95% CI: $[-0.17\%, 0.16\%]$). This difference is not significant ($p = 0.926$). This contradicts the correlational finding from the previous section, and shows that we have no evidence to reject the null hypothesis that comment moderation is not politically biased on average. Instead, this provides empirical evidence that conservative YouTube users and politicians have erroneously assumed that YouTube's moderation practices are biased against them. Rather, rightward political-lean is a proxy for other confounding variables.

H1b$_0$ is still rejected. The estimated moderation likelihood for comments under videos with *center* channels is $2.36\% \pm 0.07\%$ and with *extreme* channels is $3.57\% \pm 0.11\%$, which represents an estimated causal effect of $1.21\%^{***}$ (95% CI: $[1.04\%, 1.39\%]$). This corresponds to a 51% in-

[10] A recent study showed that such CIs are conservative estimates (Austin and Small 2014).

[11] Note that because we run four hypotheses simultaneously, we use Bonferroni correction to counteract the problem of multiple comparisons, i.e., 95% CIs are actually 98.75% CIs.

crease, which is smaller than the 71% increase from center to extreme channels we observed in the correlational tests. Regardless, we still find evidence that the magnitude of video partisanship impacts the likelihood of comment moderation. This finding may also partially explain accusations of biased content moderation, since we observe that there are a greater number of ideologically extreme right-leaning channels than similarly extreme left-leaning channels on YouTube.

H2a$_0$ is still rejected. The estimated moderation likelihood for comments under *false* videos is $3.32\% \pm 0.11\%$ and under *true* videos is $1.17\% \pm 0.05\%$, which represents an estimated causal effect of $-2.16\%^{***}$ (95% CI: $[-2.32\%, -2.02\%]$). This corresponds to a 65% decrease, which is larger than the 44% decrease from *false* to *true* videos we observed in the correlational tests, mainly because the estimated moderation likelihood for comments on true videos decreases. In sum, we find evidence that the veracity of videos affects the likelihood of moderation.

H2b$_0$ is still rejected. The estimated moderation likelihood for comments posted *before* fact-checking is $2.87 \pm 0.10\%$ and *after* fact-checking is $3.54\% \pm 0.09\%$, which represents an estimated causal effect of $0.67\%^{***}$ (95% CI: $[0.51\%, 0.83\%]$). This corresponds to a 23% increase, which is smaller than the 54% increase after fact-checking we observed in the correlational tests. This suggests that although confounding variables subsume a large part of the observed correlational difference, we still find evidence that comments are more likely to be moderated after the associated video is fact-checked.

6 Limitations & Alternative Explanations

Although we analyze our hypotheses within a relatively controlled setting, our analysis is still limited by available datasets and model specifications. In this section, we discuss the limitations and alternative explanations for our results.

Moderation Sources

One limitation of our study is our inability to determine who moderated a given comment: the video uploader, a human moderator at YouTube, an algorithm, or the commenter themselves. To address this, we use simulations to investigate how our analysis would change under varying assumptions about the fraction of comments that are removed by commenters themselves. We assume a self-moderation rate r, i.e., the remaining $1 - r$ removed comments were moderated by YouTube's systems. We randomly sample $1 - r$ of the moderated comments in our dataset while keeping the unmoderated comments the same. As shown in Figure 7, self-moderation does not change our conclusion for H1a$_0$ for a spectrum of r from 0% to 50%. Although the effect size for H2a$_0$ and H2b$_0$ fluctuate as r increases, the direction of their effects are robust. The only exception is H1b$_0$: the direction of its causal effect does not hold when $r > 20\%$.

Note that this robustness check assumes a constant user moderation rate over all moderated comments, which oversimplifies reality. The moderation behavior of video uploaders and commenters are likely correlated with unmeasured variables, e.g., video uploaders may be more likely to moderate comments that disagree with their own position, either due to direction or extremity of partisanship. Investigating when and why self-moderation happens is beyond our current capabilities, therefore we leave it for future work.

Credibility of Fact-Checking

Our credibility labels are drawn from Snopes and PolitiFact, which are both confirmed by the International Fact-Checking Network to be non-partisan, fair, and transparent (Poynter 2018). However, there are still accusations that their ratings are biased against political conservatives (Richardson 2018; Shapiro 2016; NewsBusters 2018). Although we do observe that right-leaning videos are more likely to be rated as false ($\chi^2 = 221.8^{***}$), we do not know if the political leaning actually causes this difference.

Exploring the bias of ratings from fact-checkers themselves is beyond the scope of this paper, but still, we investigate the hypothetical case where fact-checkers are systematically biased. We assume a bias b, where $b = \lambda$L represents a systematic bias against liberals and $b = \lambda$R represents bias against conservatives, and λ represents the magnitude of bias (0, non-existing; +1, slight; +2, high). We recalibrate all the veracity scores in our dataset given a value for b. For example, $b = +1$R represents a slight bias against conservatives, which we consider as a form of underrating right-leaning videos. Therefore, all conservative videos labeled as "mostly true" by the fact-checker will instead be considered true. Similarly, if $b = +2$R, then all conservative videos labeled as "half true" or "mostly true" by the fact-checker will instead be considered true.

The results of our causal models under various values of b are shown in Figure 8. H2a$_0$ is impacted the most, since it directly concerns video veracity. In contrast, the effect sizes of H1b$_0$ and H2b$_0$ fluctuate, but the direction of their effects are robust. For H1a$_0$, the result does not change with slight bias ($\lambda \leq +1$), but does change when fact-checkers are highly biased. Consider $b = +2$R, which means fact-checkers are highly biased against right-leaning videos: in the calibrated case, content moderation is also biased against right-leaning videos. *Vice versa* for $b = +2$L. Similarly, the results of H2a$_0$ also change in the same direction.

Note that **we do not support claims of bias against fact-checkers in any way**. We investigate this hypothetical scenario simply for the sake of thoroughness, i.e., to show that even if fact-checkers were slightly biased, it would not explain why comments on right-leaning videos are moderated more heavily than comments on left-leaning videos.

Alternative Thresholds & Controls

We now explore model dynamics under alternative thresholds and controls for our labels.

First, our label for right- and left-leaning video channels is based on the sign of partisanship score. However, it is conceivable that scores near zero may not indicate perceptible partisanship (Robertson et al. 2018). Therefore, we set a minimum threshold for partisanship scores, i.e., only absolute scores greater than the threshold are labeled right/left, others are considered neutral and not used for analysis. As shown in Figure 9, such thresholding has minimal impact on H1b$_0$,

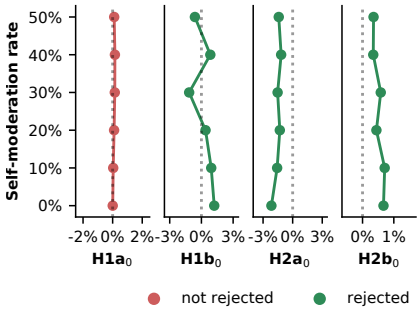

Figure 7: **Simulation of user moderation.** The effect of self moderation is minimal for $H1a_0$, $H2a_0$, and $H2b_0$, but $H1b_0$ does not hold under high rates ($r > 20\%$).

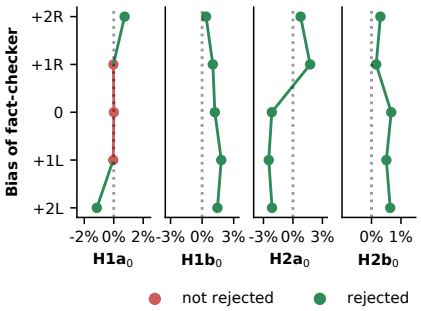

Figure 8: **Simulation of biased fact-checkers.** The effect of fact-checker bias is minimal for $H1b_0$ and $H2b_0$, and minimal for $H1a_0$ when bias is low ($\lambda \leq +1$).

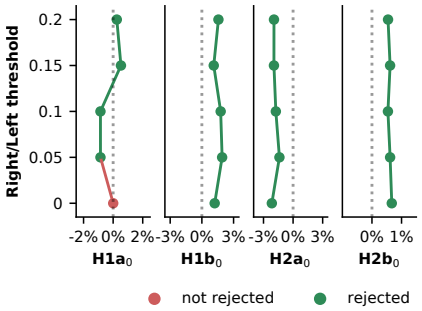

Figure 9: **Alternative $H1a_0$ (left/right) thresholds.** The effect of left/right thresholds is minimal for $H1b_0$, $H2a_0$ and $H2b_0$, but results for $H1a_0$ do not hold.

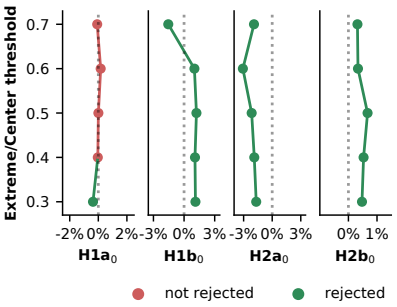

Figure 10: **Alternative $H1b_0$ (extreme/center) thresholds.** The effect of extreme/center thresholds is minimal for most hypotheses, except for $H1a_0$ and $H1b_0$.

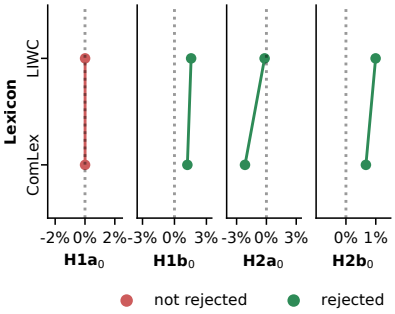

Figure 11: **Alternative linguistic controls.** The effect of alternative linguistic controls using lexicon LIWC instead of ComLex is minimal for all hypotheses.

$H2a_0$, and $H2b_0$, but does impact $H1a_0$.[12] However, since the effect fluctuates between leftward and rightward bias, the claim for "conservative bias" is still not supported overall.

Next, we investigate how alternative thresholds for extreme/center labels affects our results by replacing our original threshold 0.5 with a spectrum from 0.3 to 0.7. As shown in Figure 10, this change has minimal impact on all hypotheses with two exceptions. a) We observe leftward bias for $H1a_0$ under threshold 0.3; although this bias is statistically significant, the difference is only 0.37% which yields minimal practical impact. b) The bias flips for $H1b_0$ under threshold 0.7, but this is caused by poor model performance since such extremely partisan video channels are rare in our dataset (leading to a sample of < 1000 moderated comments).

Third, we examine an alternative set of linguistic controls using LIWC (Pennebaker, Francis, and Booth 2001; Tausczik and Pennebaker 2010). Although the ComLex lexicon is context-specific, it has not been as extensively used as LIWC. We derived five categories from LIWC: *swear, money, work, biological process,* and *punctuation*,[13], use them in place of the linguistic controls from ComLex, and rerun our model. As shown in Figure 11, the difference between using ComLex and LIWC is minimal for all hypotheses.

Concerns Regarding Causal Models

There are two main concerns when using causal models. The first is *reverse causality* (Marquis et al. 1997), which refers to the case where the direction of a causal effect may be the opposite of what is assumed, or the causal effect is a two-way relationship. Reverse causality does not apply to our study, since in our dataset the outcome variable (comment moderation) comes strictly after a video is posted, when all our hypothesized variables are already determined.

Another concern is *unmeasured confounding variables* (Robins, Rotnitzky, and Scharfstein 2000), which refers to factors that might affect the outcome and correlate with treatments but are not controlled in the model. Our controlled confounders include social engagement with YouTube videos and linguistics in user comments, which are intuitive and highly relevant given YouTube's community guidelines (YouTube 2018) and prior studies (Jiang and Wilson 2018; Shen et al. 2018; Chandrasekharan et al. 2018). However, this set is admittedly incomplete; unmeasured factors such as user characteristics, comment volume, the presence of

[12] This is partially due to the partisan bias scores of comments in our dataset not being balanced between left and right. See Supplementary Materials.

[13] Determined by a preliminary linear regression for $p < 0.001$.

"bots," etc., could still skew the results of propensity scoring models (King and Nielsen 2016). Nevertheless, the results from propensity scoring show significant improvement comparing to correlational analysis (De Choudhury et al. 2016).

Again, although causal models analyze relationships between treatments (i.e., hypotheses) and outcome (i.e., moderation), they do not explain intermediate factors. For example, it could be that extreme partisanship and high-level misinformation directly affect the attention and decision-making of algorithmic or human moderators (News 2017; Morris 2017; Chatterjee and Dave 2017; Alvarez 2018; Post 2018; Palladino 2018). Or it could be that fact-check messages draw more efforts from concerned users to flag content for moderation (Glaser 2018; McCracken 2018).

Representation & Generalization

The YouTube videos in our dataset are covered by the datasets from (Jiang and Wilson 2018) and (Robertson et al. 2018), which means they were published by identifiable entities that have web presences off YouTube, and were influential enough to draw the attention of fact-checkers. In other words, the videos in our study are higher-profile than average on YouTube. Measured by number of views, our sample of YouTube videos has a mean of 4,311,320 $\pm$ 38,942 views, which is significantly higher than the average views measured by previous studies (Cheng, Dale, and Liu 2008; Figueiredo et al. 2014; Miotto and Altmann 2014). Thus, our findings may not be representative across all videos on YouTube. That said, the vast majority of videos on YouTube receive very few views and comments, meaning they are not viable or interesting candidates for study. Instead, by focusing on high-profile videos, we present results that we believe are more relevant to the YouTube community and policymakers.

We use YouTube as a lens to investigate comment moderation as we believe that this is a vitally important endeavor at this moment in time, given the prevailing political climate. That said, we caution that our findings may not generalize beyond YouTube. Further, platform moderation policies are notoriously fickle, meaning that our findings may not generalize over time.

7 Conclusion & Implication

In this paper, we investigate how partisanship and misinformation in YouTube videos affects the likelihood of moderation among the user comments on those videos. Using a dataset of 84,068 comments posted across 258 videos, we find no evidence that comments on right-leaning videos are moderated more heavily than left-leaning videos, once measured confounding variables are controlled. Instead, the greater amount of comment moderation on right-leaning videos is explained by correspondingly higher levels of misinformation, extreme partisanship of videos, and various linguistic signals (e.g., hate speech) in comments. These moderation decisions are consistent with YouTube's community guidelines (YouTube 2018).

Our study advances the call for researchers to engage with issues of societal and political importance, especially as they pertain to a healthy web and concerns of partisan bias and free speech (Lazar et al. 2016; Chancellor and Counts 2018; Chandrasekharan et al. 2017; Epstein and Robertson 2015; Epstein et al. 2017). The major design implication stemming from our findings concerns the non-transparent deletion of comments on YouTube. Opaque moderation practices, regardless of whether they are fully or semi-automated, are a breeding ground for theories like the one we've refuted here – anti-conservative bias in moderation practices. Indeed, this is both a motivation of our study and one of the limitations of our dataset: there is no record of when, why, or by who a comment was deleted. Although moderation is absolutely a critical component of healthy social media systems (Buni and Chemaly 2016), platform providers should consider designs that are more constructive and transparent.

Towards this goal, we recommend that deleted comments be preserved and protected. That is, comments are still moderated under existing policies, but the original comment is hidden behind a notification that it has been moderated. Then, if a user or researcher is interested in what was moderated and why, they can click on the notification to view the original comment alongside the specific policy violations that caused it to be moderated. Additional meta-information could also be provided about who moderated the comment – the platform or the channel owner – and whether the comment was flagged by automated systems. This design serves two purposes. First, it would give the commenter an explanation for why their comment was deleted and provide them with feedback on how to improve their discourse. Second, because the comment and its policy violations are preserved, it provides transparency and feedback to the community at large. This transparency, in turn, may discourage public figures from making false claims about why comments were moderated, since external researchers will have the ability to fact check such claims and mitigate the damage done to the platform in terms of user trust (Woodruff et al. 2018).

The second benefit (transparency) could negate the first (feedback), however, if the user who posted the deleted comment is exposed to the community: the user may be shamed into no longer participating or worse (Klonick 2015; 2017). Instead of learning how to be civil, they may simply go elsewhere. For example, researchers found that Reddit's ban of two hate speech subreddits was effective in reducing overall hate speech usage on the site, but noted that this ban had simply "made these users (from banned subreddits) *someone else's problem*" and "likely did not make the internet safer or less hateful" (Chandrasekharan et al. 2017). To avoid this outcome, the comment should be preserved, but the offending user should be anonymized. The goal of this design is to educate, to give a human being the opportunity to learn, not to exclude. Further research is needed to investigate how this may play out in practice.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2):211–36.

Alvarez, E. 2018. Youtube ceo talks misinformation, creators and comments at sxsw. Engadget.

Arceneaux, K.; Johnson, M.; and Murphy, C. 2012. Polarized political communication, oppositional media hostility, and selective exposure. *The Journal of Politics* 74(1):174–186.

Austin, P. C., and Small, D. S. 2014. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine* 33(24):4306–4319.

Austin, P. C. 2008. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 27(12):2037–2049.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

Bickert, M.; Downs, J.; and Pickles, N. 2018. Facebook, google and twitter: Examining the content filtering practices of social media giants. House Judiciary Committee.

Buni, C., and Chemaly, S. 2016. The secret rules of the internet. The Verge.

Chancellor, S., and Counts, S. 2018. Measuring employment demand using internet search data. In *Proc. of CHI*, 122.

Chancellor, S.; Pater, J. A.; Clear, T. A.; Gilbert, E.; and De Choudhury, M. 2016. #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proc. of CSCW*.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *PACM on HCI* 1(CSCW):1–22.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *PACM on HCI* 2(CSCW):32.

Chatterjee, R., and Dave, P. 2017. Youtube set to hire more staff to review extremist video content. Independent.

Chen, L.; Ma, R.; Hannák, A.; and Wilson, C. 2018. Investigating the impact of gender on rank in resume search engines. In *Proc. of CHI*, 651.

Cheng, X.; Dale, C.; and Liu, J. 2008. Statistics and social network of youtube videos. In *Proc. of IWQoS*.

Constine, J. 2018. Facebook reveals russian troll content, shuts down 135 ira accounts. Tech Crunch.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. of CHI*, 2098–2110.

Diakopoulos, N., and Naaman, M. 2011. Towards quality discourse in online news comments. In *Proc. of CSCW*.

Dorsey, J. 2018. Twitter: Transparency and accountability. House Energy and Commerce Committee.

Epstein, R., and Robertson, R. E. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *PNAS* 112(33).

Epstein, R.; Robertson, R. E.; Lazer, D.; and Wilson, C. 2017. Suppressing the search engine manipulation effect (SEME). *PACM on HCI* 1(CSCW):1–22.

Facebook. 2018. Community standards.

Fast, E.; Vachovsky, T.; and Bernstein, M. S. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proc. of ICWSM*, 112–120.

Figueiredo, F.; Almeida, J. M.; Gonçalves, M. A.; and Benevenuto, F. 2014. On the dynamics of social media popularity: A youtube case study. *ACM ToIT* 14(4):24.

Foong, E.; Vincent, N.; Hecht, B.; and Gerber, E. M. 2018. Women (still) ask for less: Gender differences in hourly rate in an online labor marketplace. *PACM on HCI* 2(CSCW):53.

Fridkin, K.; Kenney, P. J.; and Wintersieck, A. 2015. Liar, liar, pants on fire: How fact-checking influences citizens? reactions to negative advertising. *Political Communication* 32(1):127–151.

Gibbs, S. 2017. Google says ai better than humans at scrubbing extremist youtube content. The Guardian.

Gillespie, T. 2018. There's a reason that misleading claims of bias in search and social media enjoy such traction. Medium.

Glaser, A. 2018. Youtube is adding fact-check links for videos on topics that inspire conspiracy theories. Slate.

Hannák, A.; Wagner, C.; Garcia, D.; Mislove, A.; Strohmaier, M.; and Wilson, C. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proc. of CSCW*, 1914–1933.

Hu, D.; Jiang, S.; Robertson, R. E.; and Wilson, C. 2019. Auditing the partisanship of google search snippets. In *Proc. of WWW*.

Jhaver, S.; Ghoshal, S.; Bruckman, A.; and Gilbert, E. 2018. Online harassment and content moderation: The case of blocklists. *ACM ToCHI* 25(2):1–33.

Jiang, S., and Wilson, C. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *PACM on HCI* 2(CSCW).

Jiang, S.; Chen, L.; Mislove, A.; and Wilson, C. 2018. On ridesharing competition and accessibility: Evidence from uber, lyft, and taxi. In *Proc. of WWW*.

Kamisar, B. 2018. Conservatives cry foul over controversial group's role in youtube moderation. The Hill.

King, G., and Nielsen, R. 2016. Why propensity scores should not be used for matching.

Klonick, K. 2015. Re-shaming the debate: Social norms, shame, and regulation in an internet age. *SSRN Electronic Journal*.

Klonick, K. 2017. The new governors: The people, rules, and processes governing online speech. (ID 2937985).

Lanza, S. T.; Moore, J. E.; and Butera, N. M. 2013. Drawing causal inferences using propensity scores: A practical guide for community psychologists. *American journal of community psychology* 52(3-4):380–392.

Lazar, J.; Abascal, J.; Barbosa, S.; Barksdale, J.; Friedman, B.; Grossklags, J.; Gulliksen, J.; Johnson, J.; McEwan, T.; Martínez-Normand, L.; et al. 2016. Human–computer interaction and international public policymaking: a framework for understanding and taking future actions. *Foundations and Trends® in Human–Computer Interaction* 9(2):69–149.

Levin, S. 2017. Google to hire thousands of moderators after outcry over youtube abuse videos. The Guardian.

Li, M.; Lu, Q.; and Long, Y. 2017. Are manually prepared affective lexicons really useful for sentiment analysis. In *Proc. of IJCNLP*, volume 2, 146–150.

Liao, Q. V., and Fu, W.-T. 2013. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proc. of CHI*, 2359–2368.

Loper, E., and Bird, S. 2002. Nltk: The natural language toolkit. In *Proc. of ACL Workshop*, volume 1.

Marquis, G. S.; Habicht, J.-P.; Lanata, C. F.; Black, R. E.; and Rasmussen, K. M. 1997. Association of breastfeeding and stunting in peruvian toddlers: an example of reverse causality. *International journal of epidemiology* 26(2):349–356.

Masnick, M. 2018. Internet content moderation isn't politically biased, it's just impossible to do well at scale. Techdirt.

McCracken, H. 2018. Youtube will use wikipedia to fact-check internet hoaxes. Fast Company.

Miotto, J. M., and Altmann, E. G. 2014. Predictability of extreme events in social media. *PLoS One* 9(11):e111506.

Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017. A measurement study of hate speech in social media. In *Proc. of HT*.

Morris, D. Z. 2017. Hate speech: Youtube restricts extremist videos. Fortune.

Murray, S., and Lima, C. 2018. Trump accuses social media giants of 'silencing millions of people'. Politico.

News, S. 2017. New youtube recruits to monitor online extremist propaganda 'wrong approach'. Sputnik International.

NewsBusters. 2018. Don't believe the liberal "fact-checkers"!

Newton, C. 2018. Why twitter should ignore the phony outrage over "shadow banning". The Verge.

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2(2).

Nunez, M. 2016. Former facebook workers: We routinely suppressed conservative news. Gizmodo.

Nyhan, B., and Reifler, J. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32(2):303–330.

Olteanu, A.; Castillo, C.; Boy, J.; and Varshney, K. R. 2018. The effect of extremist violence on hateful speech online. In *Proc. of ICWSM*.

Olteanu, A.; Varol, O.; and Kiciman, E. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proc. of CSCW*.

Palladino, V. 2018. Youtube to fight fake news with links to real news and context. Ars Technica.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71.

Porter, E.; Wood, T. J.; and Kirby, D. 2018. Sex trafficking, russian infiltration, birth certificates, and pedophilia: A survey experiment correcting fake news. *Journal of Experimental Political Science* 1–6.

Post, N. Y. 2018. Youtube committing $25m to fight fake news. New York Post.

Poynter. 2018. Verified signatories of the ifcn code of principles. Poynter.

RedStateMedia. 2016. Donald trump tells jake tapper he won't denounce david duke or the kkk. YouTube.

RedStateMedia. 2018. Youtube homepage for redstatemedia. YouTube.

Richardson, V. 2018. Conservative project seeks to fact-check the fact-checkers accused of liberal bias. The Washington Times.

Robertson, R. E.; Jiang, S.; Joseph, K.; Friedland, L.; Lazer, D.; and Wilson, C. 2018. Auditing partisan audience bias within google search. *PACM on HCI* 2(CSCW).

Robins, J. M.; Rotnitzky, A.; and Scharfstein, D. O. 2000. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*. 1–94.

Robinson, R. J.; Keltner, D.; Ward, A.; and Ross, L. 1995. Actual versus assumed differences in construal: "naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology* 68(3).

Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 1–23.

Shapiro, M. 2016. Running the data on politifact shows bias against conservatives.

Shen, Q.; Yoder, M.; Jo, Y.; and Rose, C. 2018. Perceptions of censorship and moderation bias in political debate forums. In *Proc. of ICWSM*.

Snyder, R. 2016. Pro-rubio super pac ad tying trump to kkk misses the mark. PolitiFact.

Solon, O. 2018. Facebook says cambridge analytica may have gained 37m more users' data. *The Guardian* 4.

Speicher, T.; Ali, M.; Venkatadri, G.; Ribeiro, F. N.; Arvanitakis, G.; Benevenuto, F.; Gummadi, K. P.; Loiseau, P.; and Mislove, A. 2018. Potential for discrimination in online targeted advertising. In *Proc. of FAT**, 5–19.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25(1).

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

Thoemmes, F. J., and Kim, E. S. 2011. A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research* 46(1):90–118.

Twitter. 2018. Rules and policies.

Usher, N. 2018. How republicans trick facebook and twitter with claims of bias. The Washington Post.

Veletsianos, G.; Kimmons, R.; Larsen, R.; Dousay, T. A.; and Lowenthal, P. R. 2018. Public comment sentiment on educational videos: Understanding the effects of presenter gender, video format, threading, and moderation on YouTube TED talk comments. *PLoS One* 13(6).

Ward, A.; Ross, L.; Reed, E.; Turiel, E.; and Brown, T. 1997. Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge* 103–135.

Wood, T., and Porter, E. 2016. The elusive backfire effect: mass attitudes? steadfast factual adherence. *Political Behavior* 1–29.

Woodruff, A.; Fox, S. E.; Rousso-Schindler, S.; and Warshaw, J. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proc. of CHI*, 656.

YouTube. 2018. Community guidelines.