# Reasoning about Political Bias in Content Moderation

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

## Abstract

Content moderation, the AI-human hybrid process of removing (toxic) content from social media to promote community health, has attracted increasing attention from lawmakers due to allegations of political bias. Hitherto, this allegation has been made based on anecdotes rather than logical reasoning and empirical evidence, which motivates us to audit its validity. In this paper, we first introduce two formal criteria to measure bias (i.e., independence and separation) and their contextual meanings in content moderation, and then use YouTube as a lens to investigate if the political leaning of a video plays a role in the moderation decision for its associated comments. Our results show that when justifiable target variables (e.g., hate speech and extremeness) are controlled with propensity scoring, the likelihood of comment moderation is equal across left- and right-leaning videos.

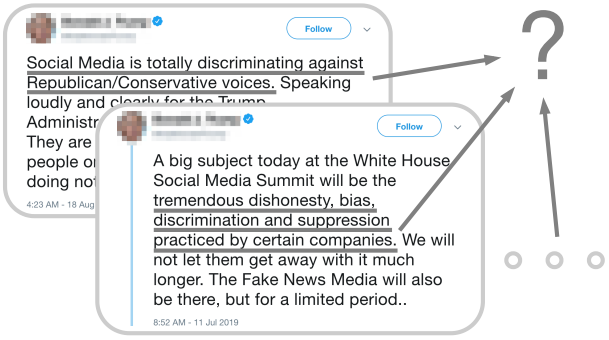



Figure 1: **Allegations of political bias** in content moderation based on anecdotes, not hard evidence.

## Bad Content Moderation, Bad!

Social media has long played host to problematic content such as partisan propaganda (Allcott and Gentzkow 2017), misinformation (Jiang and Wilson 2018), and violent hate speech (Olteanu et al. 2018). In an attempt to police this content and improve the health of their user community, social media platforms publish sets of community guidelines that explain the types of content they prohibit, and remove or hide this content from their platforms. This practice is commonly referred to as *content moderation*.

Content moderation is typically implemented as an AI-human hybrid process. To scale with the large amount of toxic content generated online, an AI filtering layer first finds potential candidates for moderation (Gibbs 2017; Sloane 2018), and then sends them to human reviewers for a final determination (Levin 2017; Gershgorn and Murphy 2017).

This content moderation process, however, has been criticized for potential bias: biased AI systems have been documented (Barocas, Hardt, and Narayanan 2019; Hutchinson and Mitchell 2019), and human moderators can bring their own biases into the moderation process (Diakopoulos and Naaman 2011). As a result, content moderation faces a backlash from ideological conservatives, who allege that social media platforms are biased against them and are censoring their views (Kamisar 2018; Usher 2018), e.g., Figure 1. These allegations have even spurred lawmakers to action, e.g., in June 2019, the "Ending Support for Internet Censorship Act" was introduced into the US Senate to limit immunity granted by Section 230 of the Communications Decency Act to "encourage providers of interactive computer services to provide content moderation that is politically neutral" (Hawley 2019).

These allegations of political bias, however, are based on anecdotes, and there is little support from logical reasoning and empirical evidence (Jiang, Robertson, and Wilson 2019; Shen et al. 2018; Shen and Rose 2019). In this paper, we conduct an audit on the validity of these allegations driven by a high-level research question:

- **Research question:** is content moderation biased?

To approach this question, we first introduce two formal criteria to measure bias and how they apply in the context of content moderation, and formulate two null hypotheses $\mathbf{H}_0^{ind}$ (for independence) and $\mathbf{H}_0^{sep}$ (for separation) under these criteria. Then, we use YouTube comment moderation as a case study to investigate a more concrete question:

- **Case study:** does the political leaning of a video play a role in the moderation decision for its comments?

Our results show: $\mathbf{H}_0^{ind}$ is rejected, i.e., comments are more likely to be moderated under right-leaning videos; $\mathbf{H}_0^{sep}$ holds, i.e., with propensity scored justifiable target variables (e.g., hate speech and extremeness), there is no significant difference in moderation likelihood across the political spectrum.

## How to Measure Bias, Really?

Recent advances in fairness research provide many criteria to measure bias, each aiming to formalize different desiderata (Barocas, Hardt, and Narayanan 2019). Most of these criteria characterize the joint or conditional probability between involved variables (e.g., decision, sensitive features), and can be approximately classified to two categories: *independence* and *separation* (Hutchinson and Mitchell 2019).

### Independence

Independence, also referred to as *demographic parity*, is a fairness criterion that requires the decision variable and the sensitive feature to be statistically independent. In the context of political bias and content moderation, an item on social media (e.g., post, comment) can be associated with its political leaning $\mathrm{P} = \{\text{left}, \text{right}\}$ and moderation decision $\mathrm{M} = \{\text{moderated}, \text{alive}\}$. This criterion requires these two variables to satisfy $\mathrm{M} \perp\!\!\!\perp \mathrm{P}$, which, given that P is a binary variable, is equivalent to:

$$\mathbb{P}\{\mathrm{M} \mid \mathrm{P} = \text{left}\} = \mathbb{P}\{\mathrm{M} \mid \mathrm{P} = \text{right}\}. \quad (1)$$

The graphic model of independence criterion is shown in Figure 2a. To allege political bias under this criterion, then, requires empirical evidence to reject (1) as the null hypothesis $\mathbf{H}_0^{ind}$ with statistical confidence.

Although this criterion is intuitive and has been applied in many studies (Robertson et al. 2018; Hu et al. 2019), its desirability is context-dependent: e.g., moderation decisions are intended to be made based on the toxicity of content, and if toxicity is unevenly distributed across the political spectrum, the pursuit for independence may be unachievable and even undesirable.

### Separation

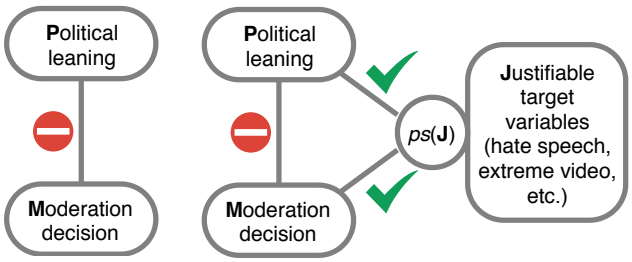
Separation, also referred to as *equalized odds*, is a type of conditional independence that allows dependence between the decision variable and the sensitive feature, but only to the extend that can be justified by target variables. For content moderation, such target variables can include hate speech, extreme videos, etc. Denoting a universe of justifiable target variables as J, this criterion requires $\mathrm{M} \perp\!\!\!\perp \mathrm{P} \mid \mathrm{J}$, which, given that P is a binary variable, is equivalent to $\forall\, \mathrm{J}$:

$$\mathbb{P}\{\mathrm{M} \mid \mathrm{P} = \text{left}, \mathrm{J}\} = \mathbb{P}\{\mathrm{M} \mid \mathrm{P} = \text{right}, \mathrm{J}\}. \quad (2)$$

This criterion is also widely adopted in previous studies, especially when the correlation between sensitive features and target variables is inherent (Thoemmes and Kim 2011; Lanza, Moore, and Butera 2013; Austin 2008).

A practical limitation of this criterion is that stable estimators of (2) requires matched observational pairs conditional on J. Therefore, as J contains more variables, matching becomes more difficult. An alternative method is to summarize all of the target variables into one scalar, i.e., $f : \mathbb{R}^{|\mathrm{J}|} \to \mathbb{R}$. A particular example of $f$ is *propensity scoring* defined as: $ps(\mathrm{J}) := \mathbb{P}\{\mathrm{P} = \text{left (or right)} \mid \mathrm{J}\}$ (Rosenbaum and Rubin 1983). It is proven that if (2) holds and $\mathbb{P}\{\mathrm{P} \mid \mathrm{J}\} \in (0, 1)$, then $\forall\, ps(\mathrm{J}), \mathbb{P}\{\mathrm{P} \mid ps(\mathrm{J})\} \in (0, 1)$ and:

$$\mathbb{P}\{\mathrm{M} \mid \mathrm{P} = \text{left}, ps(\mathrm{J})\} = \mathbb{P}\{\mathrm{M} \mid \mathrm{P} = \text{right}, ps(\mathrm{J})\}. \quad (3)$$




(a) **Independence.** 1st null hypothesis $\mathbf{H}_0^{ind}$: $\mathrm{M} \perp\!\!\!\perp \mathrm{P}$.

(b) **Separation.** Propensity scoring function $ps(\mathrm{J})$ is used to summarize J to a scala, hence 2nd null hypothesis $\mathbf{H}_0^{sep}$: $\mathrm{M} \perp\!\!\!\perp \mathrm{P} \mid ps(\mathrm{J})$.

Figure 2: **Graph models of fairness criteria.** These criteria characterize the joint or conditional distribution of political leaning P (sensitive feature), moderation decision M (decision variable) and justifiable target variables J.

The graphic model of propensity scored separation criterion is shown in Figure 2b. To allege political bias under this criterion, then, requires empirical evidence to reject (3) as the null hypothesis $\mathbf{H}_0^{sep}$ with statistical confidence.

## Is YouTube Biased, for Example?

YouTube is one of the major social media platforms that faces allegations of politically biased content moderation, and it practices moderation at different content levels, e.g., videos, channels, users, and comments (YouTube 2018a). Here, we use YouTube as a lens to investigate if the political leaning of a video plays a role in the moderation decision for its associated comments.[1]

### Data

The YouTube data we use contain 84,068 comments posted on 258 political videos,[2] labeled with involved variables.

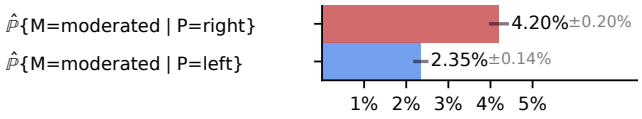
The **moderation decision** for each comment is labeled by comparing two snapshots of the dataset: the first collected in January 2018 (Jiang and Wilson 2018), and the second in June 2018 (Jiang, Robertson, and Wilson 2019). Disappeared comments within this time range are labeled as *moderated*, and the others are labeled as *alive*.

The **political leaning** of the video under which the comment was posted is labeled from another dataset (Robertson et al. 2018), where all political entities on the web are assigned an ideological score $i \in [-1, 1]$ (left to right). We link a video's publisher to its political entity, and use the sign of the ideological score as its political leaning *left* or *right*.

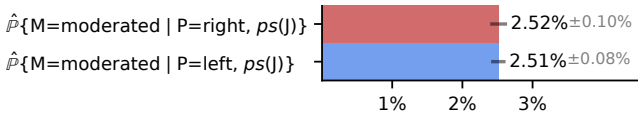
**Linguistic signals** in comments are used as the first set of target variables, as the text content of comments is the primary focus of the moderation system (YouTube 2018c). We use an existing lexicon *Comlex* to map the text content to 8 binary variables: *swear* (including hate speech, e.g., the n-word), *laugh* (e.g., "haha"), *emoji*, *fake* (e.g., "lie"), *ad-*

[1] These results, although under a different frame, are also reported in (Jiang, Robertson, and Wilson 2019).

[2] Available at: https://moderation.shanjiang.me

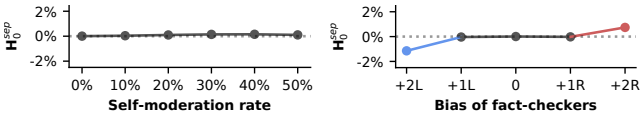



(a) $\mathbf{H}_0^{ind}$ **is rejected.** There is significant difference between comment moderation probability under left- and right- leaning videos.



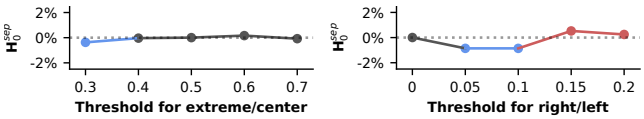

(b) $\mathbf{H}_0^{sep}$ **holds.** There is no significant difference between comment moderation probability under left- and right- leaning videos with propensity scored justifiable variables.

Figure 3: **Estimated moderation probability.** The (conditional) moderation probability for comments with corresponding confidence intervals are estimated to test the independence and separation hypotheses.




(a) **Self-moderation.** The effect of self-moderation is minimal.

(b) **Biased fact-checkers.** Slight bias does not change results.




(c) **Thresholding extremeness.** Varying thresholds for extremeness has mostly minimal effect.

(d) **Thresholding political leaning.** Results fluctuate on both the left and right ends.

Figure 4: **Robustness of $\mathbf{H}_0^{sep}$.** Potential scenarios are simulated to check the robustness of our results. The results are mostly stable except a few cases where the results fluctuate on both left and right.

*ministration* (e.g., "mayor"), *American* (e.g., "nyc", "texas"), *nation* (e.g., "mexico"), and *personal* (e.g., "your").

The **social engagement** of a video is also considered as target variables, e.g., a video with a high dislike rate attracts more flaggers and more attention from moderators. This includes three variables: *views*, *likes*, and *dislikes* of the video.

We also consider the **extremeness** of a video, as extreme videos are more likely to call for violence or spread conspiracy theories (YouTube 2018b). This is labeled using the same dataset as the political leaning variable. We label a video *extreme* if $|i| > 0.5$ and *center* otherwise.

**Misinformation** related features are another set of target variables we control. Misinformation might play a role in content moderation as social media companies have recently established collaborations with fact-checkers (e.g., Snopes, PolitiFact) (Glaser 2018). This contains two binary variables: if a video contains misinformation or not (as judged by Snopes or PolitiFact), and if a comment is posted before or after the corresponding fact-check.

## Results

Independence $\mathbf{H}_0^{ind}$ measures only moderation decision and political leaning variables. We estimate the empirical probability of the two variable and confidence intervals for binomial proportions. As shown in Figure 3a, there is significant difference between $\hat{\mathbb{P}}\{\mathrm{M} = \text{moderated} \mid \mathrm{P} = \text{left}, \mathrm{J}\}$ and $\hat{\mathbb{P}}\{\mathrm{M} = \text{moderated} \mid \mathrm{P} = \text{right}, \mathrm{J}\}$, therefore the independence hypothesis $\mathbf{H}_0^{ind}$ **is rejected.**

However, as we discuss above, the independence criterion has a fatal limitation in this context because there are strong correlations between political leaning and justifiable target variables, e.g., comments under right-leaning videos contain significantly more swear words (Pearson $\chi^2 = 671.2^{***}$),[3] indicating increased likelihood of hateful content, which is the primary trigger for content moderation. Right-leaning videos also have significantly more dislikes (Mann-Whitney $U = 4.08 \times 10^{8***}$) than left-leaning ones, providing an alternative explanation that the higher dislike rate may result in more flagged comments, thus increased moderation.

Therefore, our main focus is on investigating if the difference in moderation probability can be justified by target variables, i.e., $\mathbf{H}_0^{sep}$. We estimate propensity scores $ps(\mathrm{J})$ by logistic regressions, use them to match two-nearest neighbors from our observations, and then compute the conditional probability given matched propensity scores. As shown in Figure 3b, there is no significant difference between $\hat{\mathbb{P}}\{\mathrm{M} = \text{moderated} \mid \mathrm{P} = \text{left}, ps(\mathrm{J})\}$ and $\hat{\mathbb{P}}\{\mathrm{M} = \text{moderated} \mid \mathrm{P} = \text{right}, ps(\mathrm{J})\}$, therefore no evidence to reject the second hypothesis, i.e., $\mathbf{H}_0^{sep}$ **holds.**

Overall, our results show that although comments are more likely to be moderated under right-leaning videos, this difference is well-justified, i.e., once our measured target variables are balanced, there is no significant difference in moderation likelihood across the political spectrum.

## Robustness

Conclusions made from observational data are often questionable due to alternative explanations. Therefore, we conduct additional experiments to check the robustness of $\mathbf{H}_0^{sep}$, including self-moderation instead of moderation by the platform (Figure 4a), bias in the ratings provided by fact-checkers (Figure 4b), varying the threshold to label extremeness (Figure 4c), and labeling political leaning only when $|i|$ exceeding a certain threshold (Figure 4d). Due to space limits, we omit details on the implementation and discussion of these experiments. Interested readers can refer to our original paper (Jiang, Robertson, and Wilson 2019).

In short, these experiments show that $\mathbf{H}_0^{sep}$ in general holds when the simulated scenario is moderate, but can be rejected under extreme cases. However, under these scenarios, the results fluctuate on both the left and right ends, i.e., bias against both left and right political leanings. Therefore, the allegation of bias in YouTube comment moderation is still not supported.

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

## It's Complicated.

By using YouTube content moderation as a lens, our results show that the allegation of biased content moderation is supported by the intuitive (yet out-of-context) independence criterion, however, it is not supported by the separation criterion, where we justify moderation decisions with other target variables. Interestingly, research on alleged political bias often reaches similar conclusions: (Bakshy, Messing, and Adamic 2015) show that the allegation of biased newsfeed on Facebook was due more to homophily than algorithmic curation; (Robertson et al. 2018) show that the allegation of biased search results from Google was dependent largely on the input query instead of the self-reported ideology of the user.

The goal of this research is twofold. First, we call for transparency in content moderation practices. The opaque nature of this process can breed conspiracy theories such as the one we investigated in this paper. Further, these allegations are challenging to validate, as neither researchers, nor critics, can access removed data that underpin moderation decisions. Therefore, we recommend that moderated content be preserved and protected.

Second, our work fits in a broader scope of understanding fairness, discrimination, neutrality, and bias in algorithm-mediated systems (Baeza-Yates 2016; Sandvig et al. 2014). We introduce recently proposed theories of fairness measures (Barocas, Hardt, and Narayanan 2019; Hutchinson and Mitchell 2019) to an existing body of empirical work on auditing political bias on the web (Jiang, Robertson, and Wilson 2019; Robertson et al. 2019; Hu et al. 2019; Ali et al. 2019; Jiang, Martin, and Wilson 2019; Robertson et al. 2018). We hope this work can foster a healthier, contextual, and dialectical discussion of political bias and social media at large.

## Acknowledgments

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

Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *PACM on HCI* 4(CSCW).

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

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

Baeza-Yates, R. 2016. Data and algorithmic bias in the web. In *Proc. of WebSci*.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

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

Gershgorn, D., and Murphy, M. 2017. Facebook is hiring more people to moderate content than twitter has at its entire company. Quartz.

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

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

Hawley, J. 2019. Ending support for internet censorship act.

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

Hutchinson, B., and Mitchell, M. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proc. of FAT**.

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.; Martin, J.; and Wilson, C. 2019. Who's the guinea pig?: Investigating online a/b/n tests in-the-wild. In *Proc. of FAT**.

Jiang, S.; Robertson, R. E.; and Wilson, C. 2019. Bias misperceived: The role of partisanship and misinformation in youtube comment moderation. In *Proc. of ICWSM*.

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

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).

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

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

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).

Robertson, R. E.; Jiang, S.; Lazer, D.; and Wilson, C. 2019. Auditing autocomplete: Suggestion networks and recursive algorithm interrogation. In *Proc. of WebSci*.

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

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination*.

Shen, Q., and Rose, C. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in reddit?s quarantine policy. In *Proc. of ALW3 ACL*.

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

Sloane, G. 2018. Facebook pursues ai in bid to id harmful content. AdAge.

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

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

YouTube. 2018a. Community guidelines.

YouTube. 2018b. Harassment and cyberbullying policy.

YouTube. 2018c. Hate speech policy.