

Reasoning about Political Bias in Content Moderation

Shan Jiang, Ronald E. Robertson, and Christo Wilson

based on the ICWSM 2019 paper

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



Northeastern University

Background: what is content moderation?

Reasoning about Political Bias in **Content Moderation**



What is it?

Background: normal comments on social media

This is so misleading...

Don't hate, just vote!



Background: inappropriate comments

This is so misleading...

I just realized something, There is a n*gger shitting in the whitehouse.

Don't hate, just vote!

Shut your gay f*g slut whore skank mouth.



Background: community guidelines

This is so misleading...

I just realized something, There is a n*gger shitting in the whitehouse.

Don't hate, just vote!

Shut your gay f*g slut whore skank mouth.



Hateful content [1]

Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes:

- Ethnicity
- Gender Identity and Expression
- Sex/Gender
- Sexual Orientation



[1] *YouTube Community Guidelines*, <https://www.youtube.com/about/policies/#community-guidelines>

Background: removal, ban, etc.

This is so misleading...

I just realized something, There is a n*gger sitting in the whitehouse.

Don't hate, just vote!

Shut your gay f*g slut whore skank mouth.

Hateful content [1]

Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes:

- Ethnicity
- Gender Identity and Expression
- Sex/Gender
- Sexual Orientation



[1] *YouTube Community Guidelines*, <https://www.youtube.com/about/policies/#community-guidelines>

Background: how is it related political bias?

Reasoning about Political Bias in Content Moderation

How is it related to political bias?

Background: allegations of political bias

Social Media is totally discriminating against Republican/ Conservative voices. Speaking loudly and clearly for the Trump Administration, we won't let that happen. They are closing down the opinions of many people on the RIGHT, while at the same time doing nothing to others...

(18 Aug 2018)



Background: more allegations of political bias

Social Media is totally discriminating against Republican/

Co

Tru

clo

wh

(18

A big subject today at the White House Social Media Summit will be the tremendous dishonesty, bias, discrimination and suppression practiced by certain companies. We will not let them get away with it much longer, The Fake News Media will also be there, but for a limited period...

(11 Jul 2019)



Background: law markers in action

Social Media is totally discriminating against Republican/
Co
Tru
clo
wh
(18

A big subject today at the White House Social Media
Su
dis
co
lon
lin
(1

US - S.1914: Ending Support for Internet Censorship Act

This bill prohibits a large social media company from moderating information on its platform from a politically biased standpoint [2].

(19 Jun 2019)

CONGRESS.GOV

[2] *Ending Support for Internet Censorship Act*, <https://www.congress.gov/bill/116th-congress/senate-bill/1914>

Background: motivation

- Allegations based on **anecdotes**.
- No support from **empirical evidence**.
- *e.g.*, tendency to overestimate bias based on personal, anecdotal experience [3].



- Investigate these allegations via **scientific methods**.

[3] Shen et al., *Perceptions of Censorship and Moderation Bias in Political Debate Forums*, ICWSM 2018

Background: research question

- Is content moderation biased?

Background: case study

- Is content moderation biased?



different norms across communities [4]

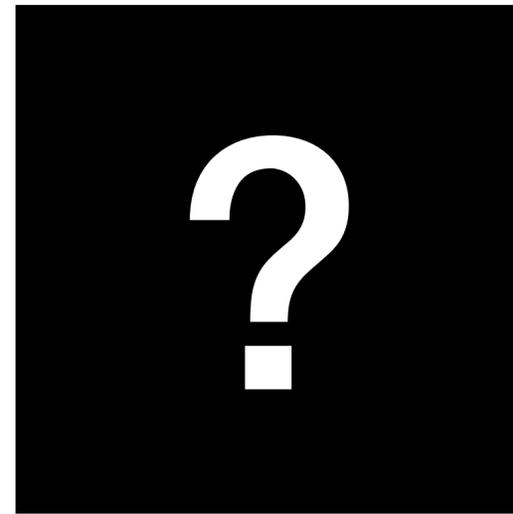
- Does the political leaning of a YouTube video play a role in the moderation decision for its comments?



[4] Chandrasekharan et al., *The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales*, CSCW 2018

Method: how to study this case?

YouTube's content moderation process



AI-human hybrid decision making

Method: external audits of black-box models

YouTube's content moderation process



AI-human hybrid decision making

Method: sensitive features and decision variable

Comments



Moderation?

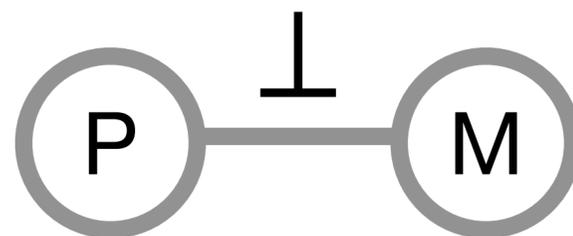
Sensitive features
e.g., political leaning
 $P = \{ \text{left, right} \}$

Moderation decision
 $M = \{ \text{moderated, alive} \}$

Method: fairness criterion - *independence*

$$\mathbb{P}\{M \mid P = \text{left}\} = \mathbb{P}\{M \mid P = \text{right}\}$$

Comments



Moderation?

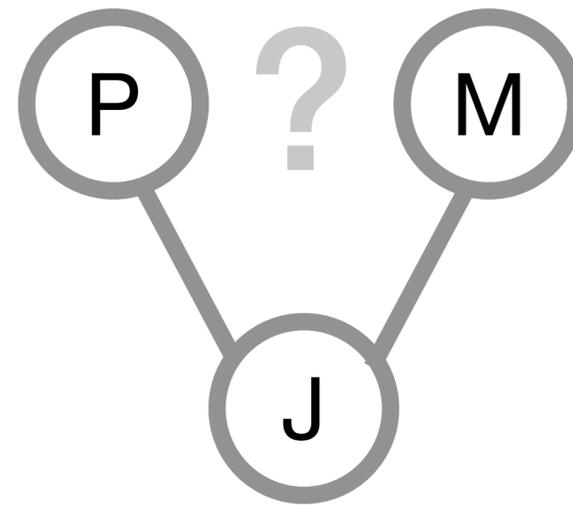
Sensitive features
e.g., political leaning
 $P = \{ \text{left}, \text{right} \}$

Moderation decision
 $M = \{ \text{moderated}, \text{alive} \}$

Method: missing justification variables?

Comments

Sensitive features
e.g., political leaning
 $P = \{ \text{left, right} \}$



Moderation?

Moderation decision
 $M = \{ \text{moderated, alive} \}$

Justifiable variables

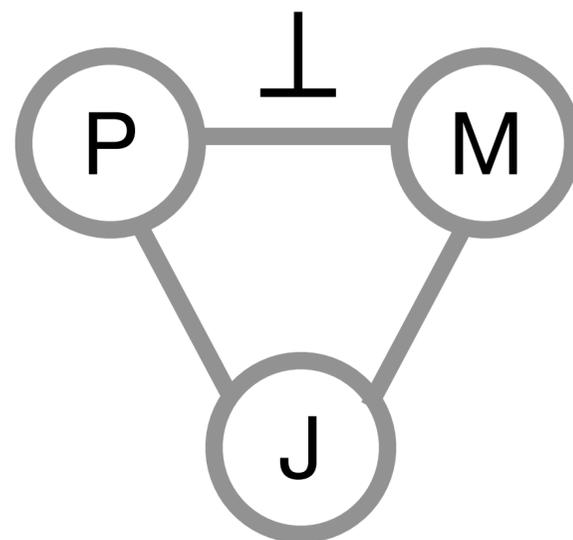
Hate speech, extreme video, etc.

Method: fairness criterion - *separation*

$$\mathbb{P}\{M \mid P = \text{left}, J\} = \mathbb{P}\{M \mid P = \text{right}, J\}$$

Comments

Sensitive features
e.g., political leaning
 $P = \{ \text{left}, \text{right} \}$



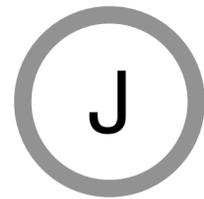
Moderation?

Moderation decision
 $M = \{ \text{moderated}, \text{alive} \}$

Justifiable variables

Hate speech, extreme video, etc.

Method: practical concern



- J is correlated with P.
- How to estimate $\mathbb{P}\{M \mid P, J\}$?



- Regression?
- Multicollinearity.

Method: *propensity score*

J

- J is correlated with P.
- How to estimate $\mathbb{P}\{M | P, J\}$?



$ps(J)$

- Propensity score.
- $ps(J) = \mathbb{P}\{P = p | J\}$.
- Estimating $\mathbb{P}\{M | P, ps(J)\}$ instead.

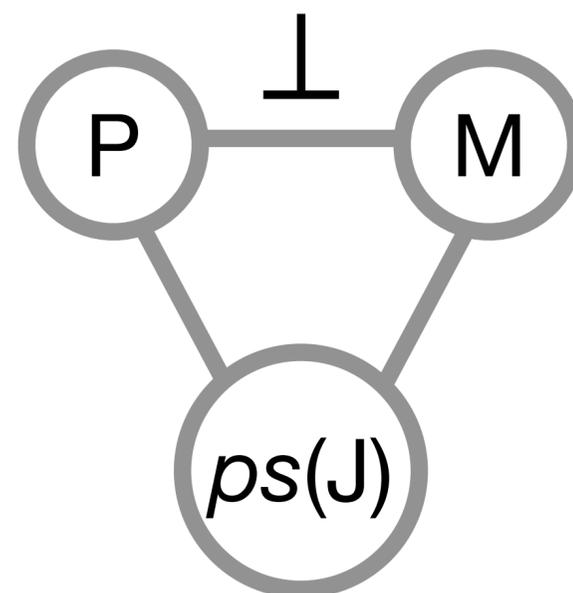


Method: realization of separation

$$\mathbb{P}\{M \mid P = \text{left}, ps(J)\} = \mathbb{P}\{M \mid P = \text{right}, ps(J)\} \quad [5]$$

Comments

Sensitive features
e.g., political leaning
 $P = \{ \text{left}, \text{right} \}$



Moderation?

Moderation decision
 $M = \{ \text{moderated}, \text{alive} \}$

Propensity score on justifiable variables
Hate speech, extreme video, etc.

[5] Rosenbaum and Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika 1983

Method: hypotheses

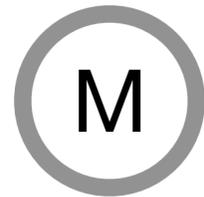
- **H₀** (independence)

$$\mathbb{P}\{M \mid P = \text{left}\} = \mathbb{P}\{M \mid P = \text{right}\}$$

- **H₀** (separation)

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

Results: dataset - M



- 84,068 comments on 258 YouTube videos.
- Two snapshots: Jan 2018 & Jun 2018.
- Missing comments are moderated.

Results: dataset - P

M

- 84,068 comments on 258 YouTube videos.
- Two snapshots: Jan 2018 & Jun 2018.
- Missing comments are moderated.

P

- Audience distribution of Democrats & Republicans.
- Left if more Democrats than Republicans.
- Right if more Republicans than Democrats.

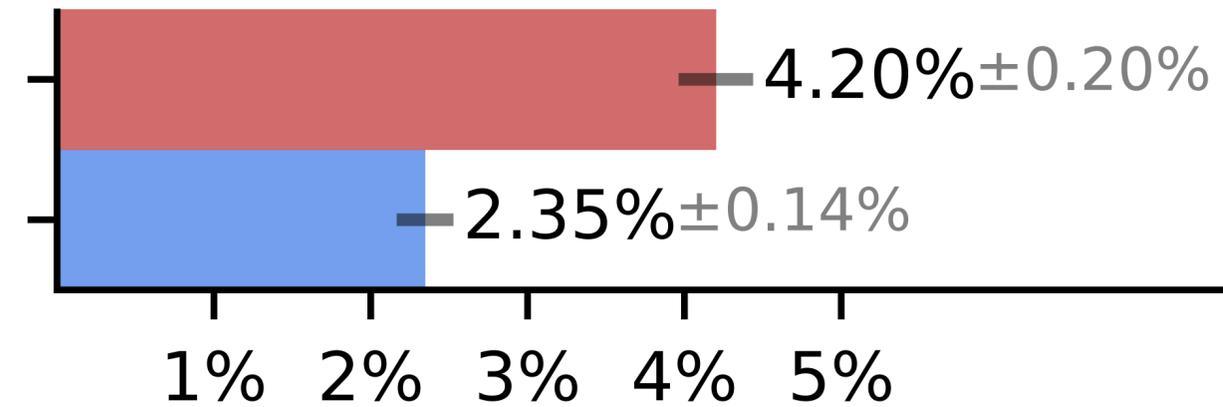
(More details in [6] and [7].)

[6] Robertson et al., *Auditing Partisan Audience Bias within Google Search*, CSCW 2018

[7] Jiang et al., *Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation*, ICWSM 2019

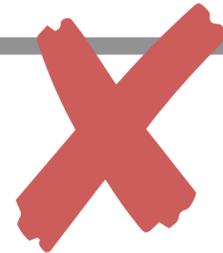
Results: independence hypothesis

$$\hat{P}\{M=\text{moderated} \mid P=\text{right}\}$$

$$\hat{P}\{M=\text{moderated} \mid P=\text{left}\}$$


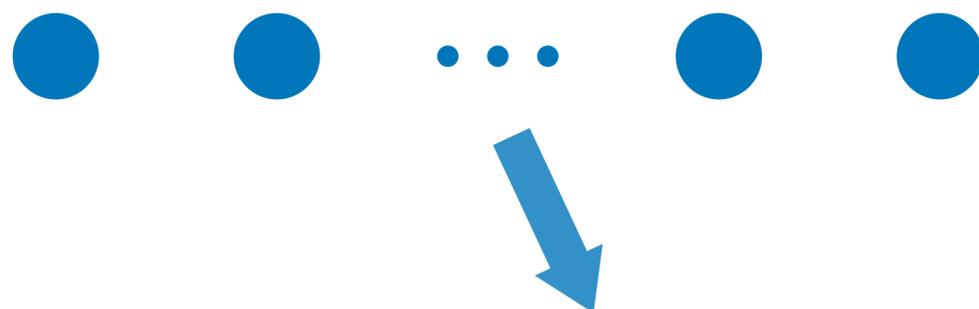
- **H₀** (independence)

$$P\{M \mid P = \text{left}\} = P\{M \mid P = \text{right}\}$$



Not Final Conclusion

Results: dataset - J (linguistic signals)



Linguistic signals of the comment, which is intuitively the most important feature for moderation decision.

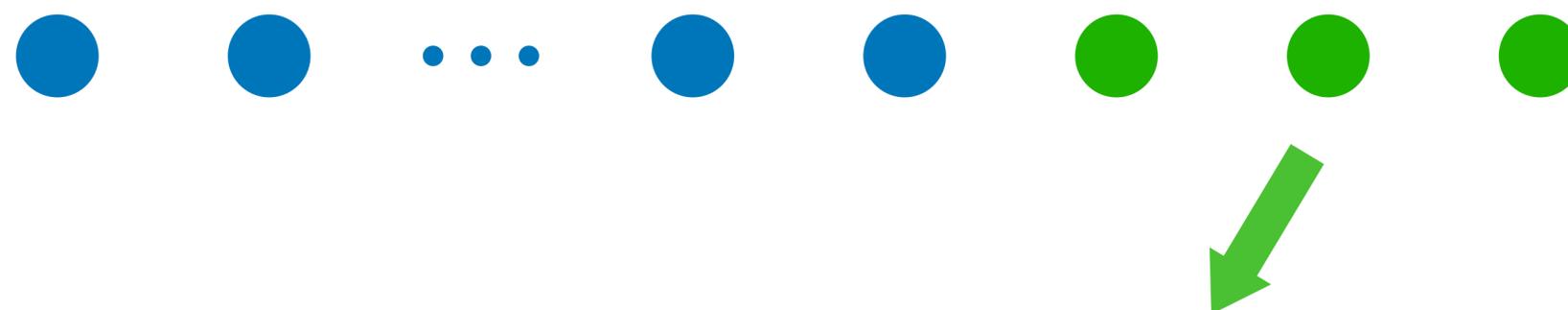
Estimated by lexicon-based frequency.
(8 features, **swear**, **laughter**, **emoji**, etc.)



Why not embeddings?
Embeddings learned from context already “embed” certain bias in the representations [8],
lexicon-based approach is more justifiable.

[8] Caliskan et al., *Semantics derived automatically from language corpora contain human-like biases*, 2017

Results: dataset - J (social engagement)

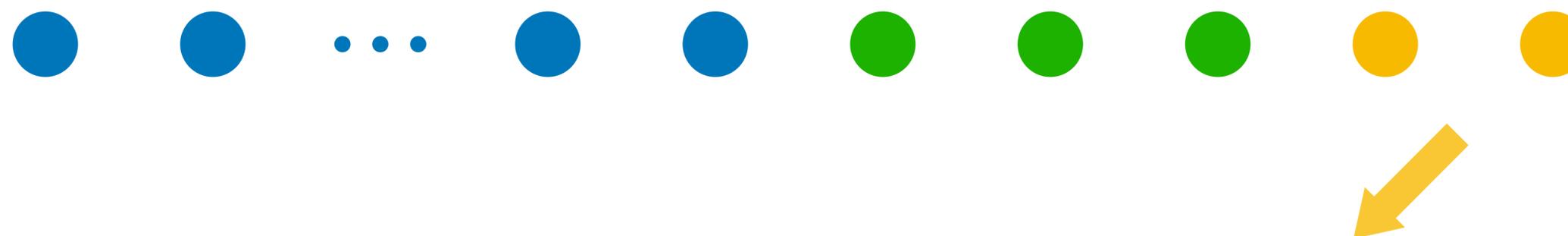


Social engagement of the video, which is correlated with the intuitive attentions from the platform.

Obtained from YouTube API.

(3 features, **views**, **likes** and **dislikes**.)

Results: dataset - J (misinformation)

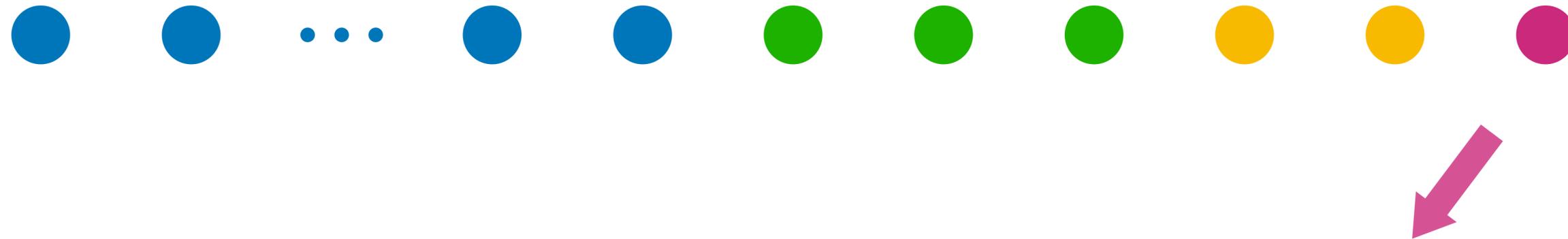


Misinformation in the video, platforms' efforts to fight misinformation [9].

Obtained by linking videos to fact-checks.
(2 features, **veracity** of the video and the comment posted **before/after** the factcheck.)

[9] Glaser, *Youtube is adding fact-check links for videos on topics that inspire conspiracy theories*, 2018

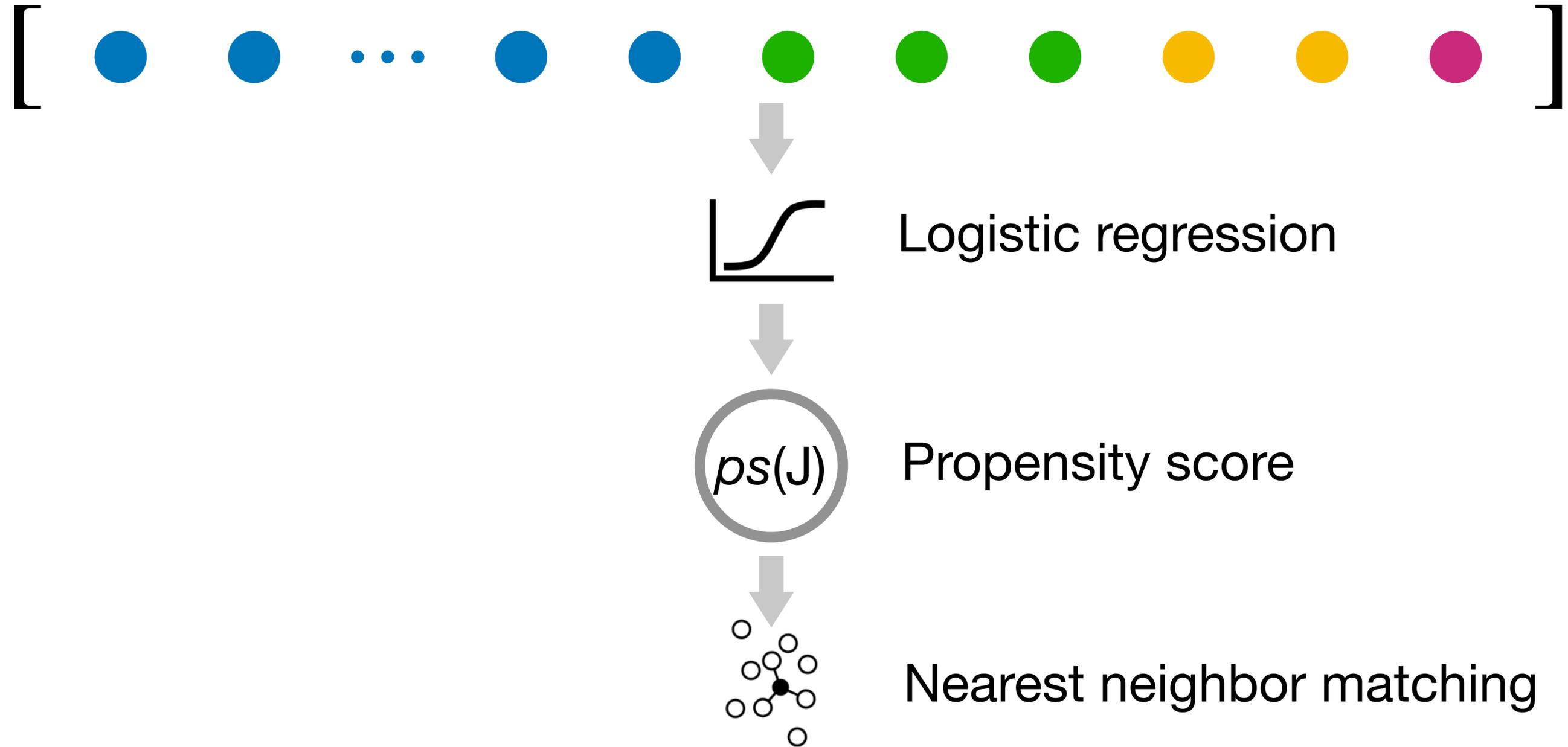
Results: dataset - J (extremeness)



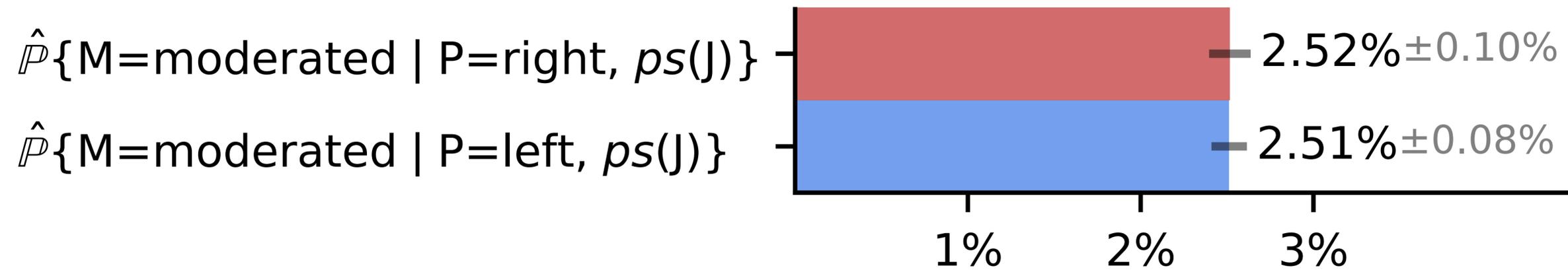
Extremeness of an outlet,
e.g., reasonable to compare New York
Times with Fox News, not with InfoWars.

Estimated from audience distribution.
(1 feature, **extremeness**)

Results: estimating propensity score



Results: separation hypothesis



- **H₀** (separation)

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



Final Conclusion

Results: robustness check

Observational empirical research:

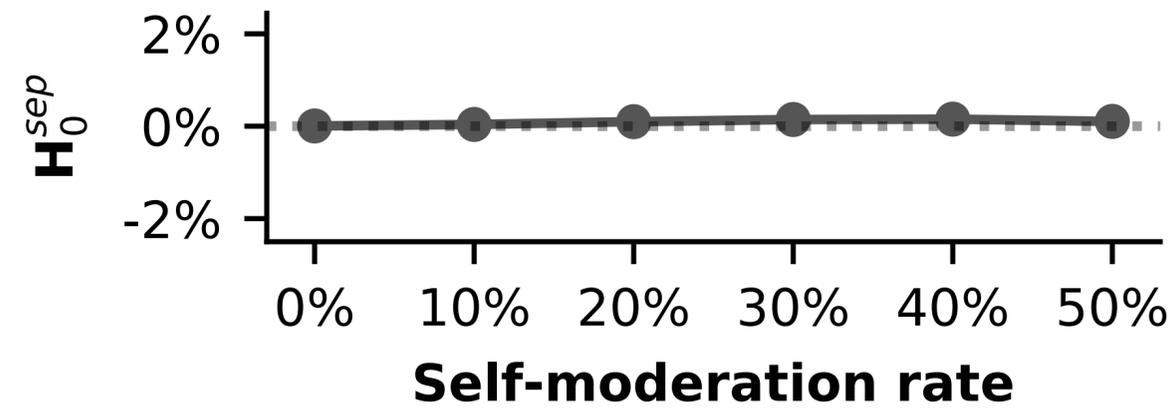
- Limitations of data collection & feature curation.
- Limitations of methods, *e.g.*, propensity score is criticized [10].



- Explore alternative cases as robustness check.

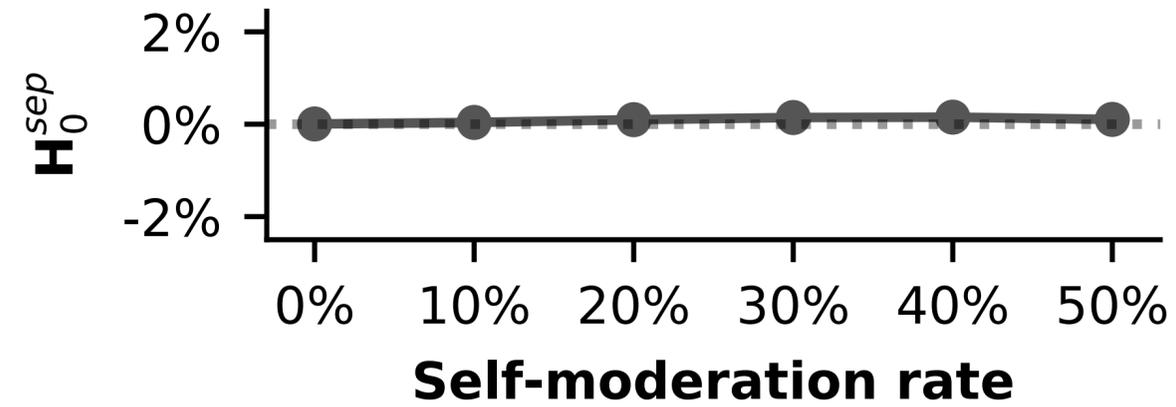
[10] King and Nielsen, *Why Propensity Scores Should Not Be Used for Matching*, 2018

Results: potential self-moderation?

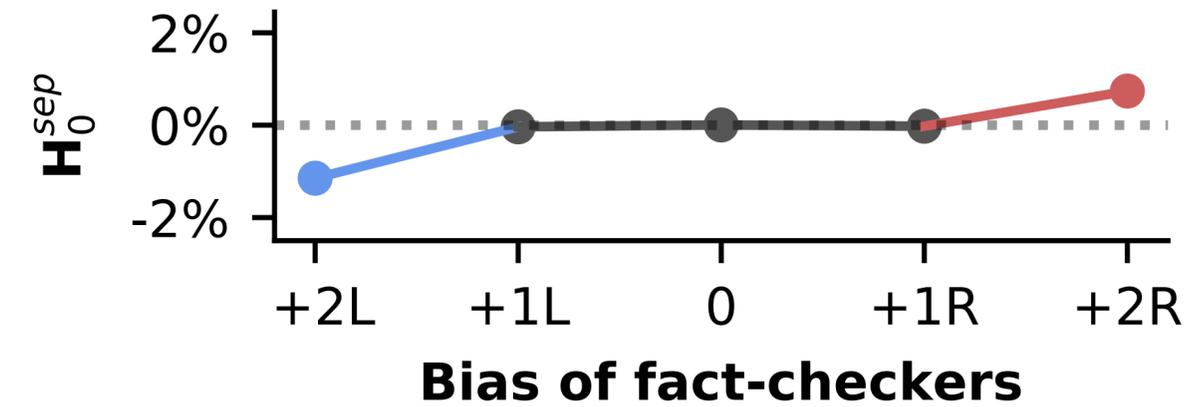


- Some comments are moderated by user, instead of the platform?

Results: biased fact-checkers?

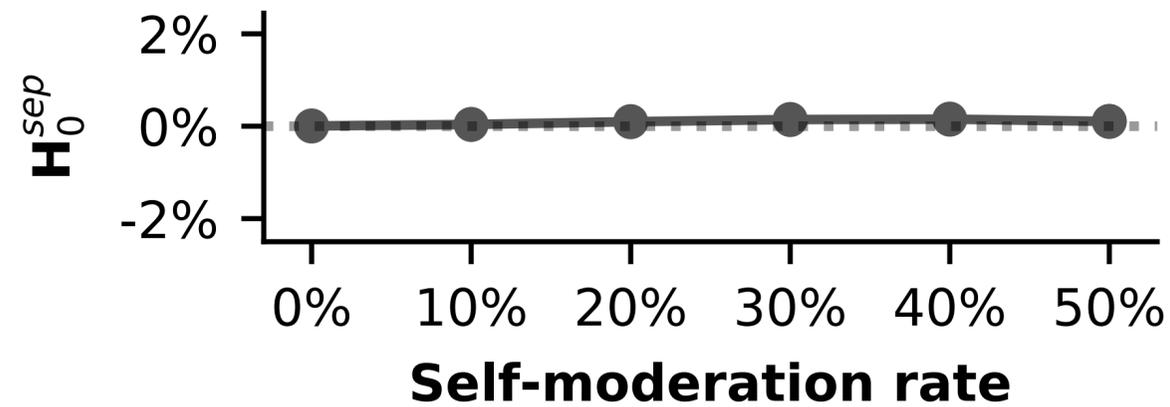


- Some comments are moderated by user, instead of the platform?

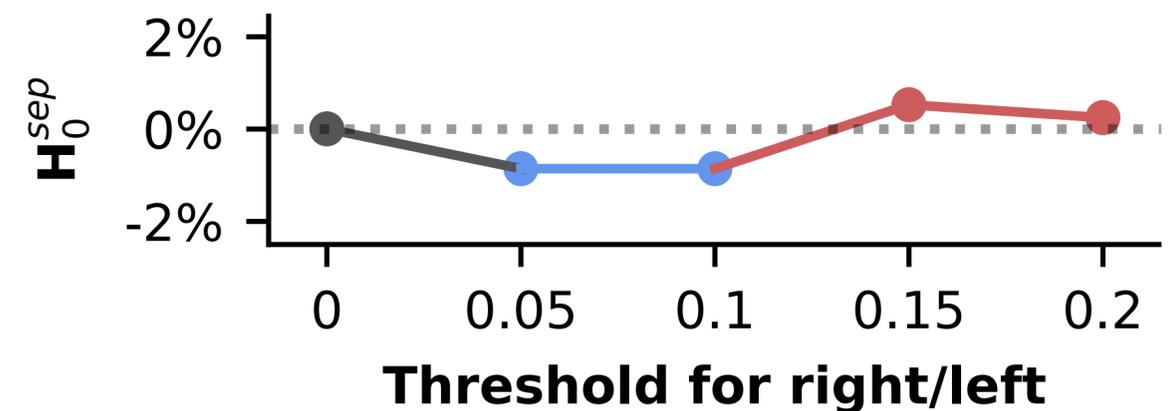


- Fact-checker themselves are biased? (over-/under-rating)

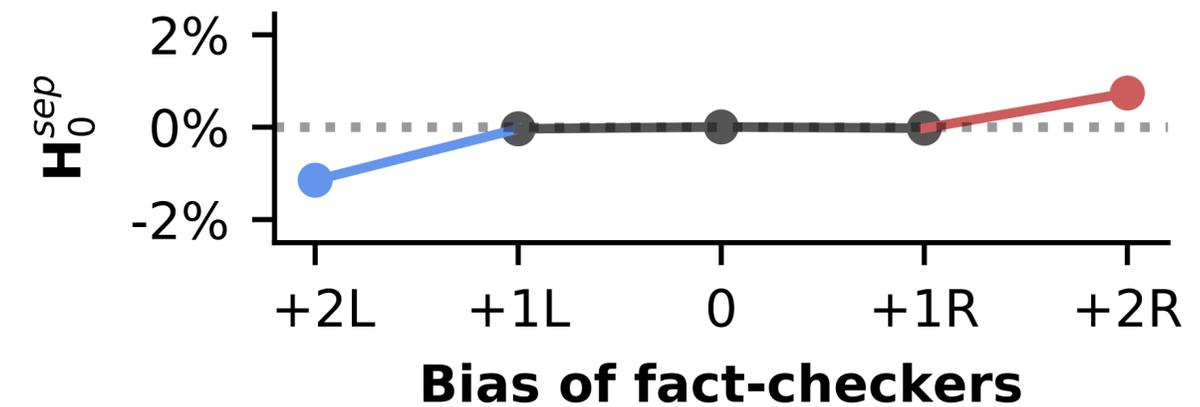
Results: changing thresholds?



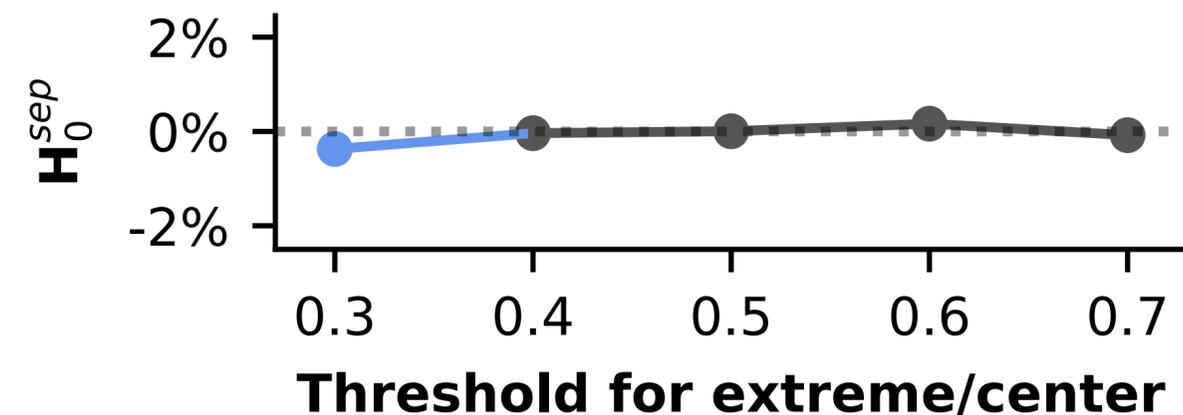
- Some comments are moderated by user, instead of the platform?



- Omit the videos with only very slight political bias?



- Fact-checker themselves are biased? (over-/under-rating)



- Change the threshold of how we define extreme?

Results: summary

- More & detailed robustness checks can be found in our original paper.
- In sum, under reasonable alternative cases, no evidence to reject:

- **H₀** (separation)

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



Discussion: what is answered?

- What is? ← Empirical question.

Discussion: what is missing?

- What is? ← Empirical question.
- What should be? ← Normative question.
- *e.g.*, what is justifiable?

Discussion: **takeaway**

- What is? ← Empirical question.
- What should be? ← Normative question.
- *e.g.*, what is justifiable?

(perspective, not a definitive answer.)

data & code available at: moderation.shanjiang.me

Thanks! Questions?

Shan Jiang
email: sjiang@ccs.neu.edu



Northeastern University