YouTube, The Great Radicalizer?

Auditing and Mitigating Ideological Biases in YouTube Recommendations

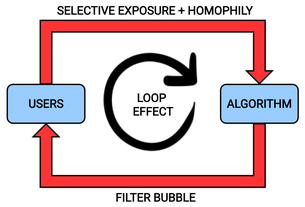
Muhammad Haroon[1], Anshuman Chhabra[1], Xin Liu[1], Prasant Mohapatra[1], Zubair Shafiq[1], Magdalena Wojcieszak[2]

[1]Department of Computer Science, University of California, Davis
[2]Department of Communication, University of California, Davis

YouTubeAudit.com

Project Statement

- Recommendations algorithms of social media platforms are often criticized for placing users in "rabbit holes" of ideologically biased content.
- While there are many potential sources of bias in a social recommendation system that factors interact in a closed-loop, algorithms are the least well-understood, having inconsistent prior evidence.



- Concerning YouTube's recommendation algorithm, two key questions remain understudied.
  - To what extent do human biases reflected in a user's watch history drive ideologically biased recommendations?
  - How to design interventions that can effectively reduce ideological bias and radicalization?

Project Goals

- Conduct a systematic audit of YouTube using sock puppets to determine the presence of ideological bias and radicalization.
- Design and evaluate a bottom-up intervention to minimize said bias.

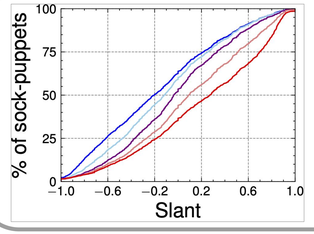
Methodology

- We conduct the audit by training and testing sock puppets: automated browser instances that mimic YouTube users by watching videos and gathering recommendations.
- The sock puppets are trained on videos from the five ideology categories: left, center-left, center, center-right, and right. The ideology of videos is estimated using Barberá's approach.

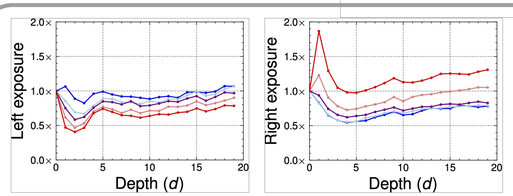
Assign ideology to sock puppet | Watch videos of assigned ideology | Collect homepage and up-next recommendations

- For mitigation, we design a Reinforcement Learning (RL) based intervention approach that minimizes ideological bias by systematically injecting ideologically diverse videos in the user's watch history to manipulate their recommendations.

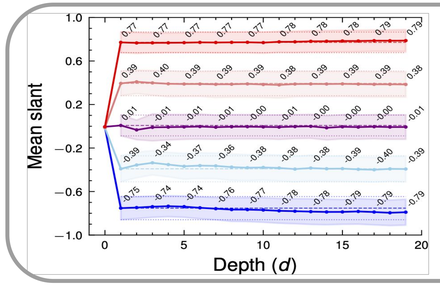
Audit



- The plot shows the distributions of the slants estimated for the recommendations of all five sock puppets.
- From left-to-right, the medians for left, center-left, center, center-right, and right sock puppets are -0.20, -0.12, -0.04, 0.09, and 0.27 respectively. This shows that the recommendations are biased towards the corresponding ideology of the sock puppet.
- The differences between these distributions are statistically significant ($p < 0.05$).



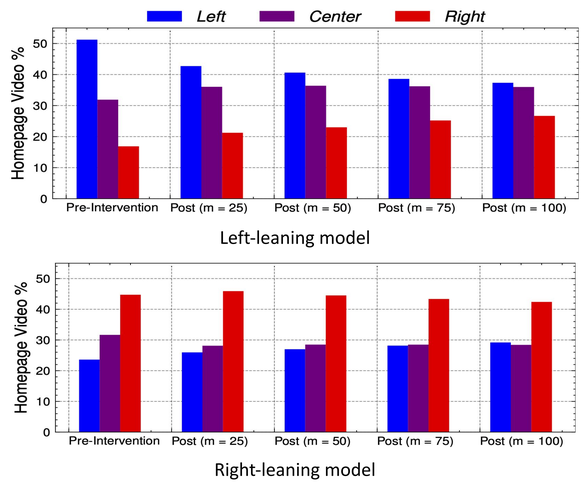
- Measuring the increase in the percentage of left and right videos as we traverse the up-next recommendations, we notice an incline in exposure as we go deeper.
- This increases the odds of the user continuing to watch ideologically biased videos.



- Plotting the mean and standard deviations of the slant of the distributions clearly shows that the mean slant of the videos for the left and the right sock puppets gradually becomes more extreme as we traverse the trail.
- For the left ideology, it rises from -0.75 to -0.79 and for the
- right ideology it rises from +0.77 to +0.79.
- Although these changes are not drastic, they do suggest that the videos deeper in the recommendation trail are more radical.

Left *Center-left* *Center* *Center-right* *Right*

Intervention



Left-leaning model

Right-leaning model

- Plot shows distribution of homepage recommendations pre-intervention and after intervention using *m* injections.
- For the left-leaning model (top), the % of left videos decreases from 51.23% to 37.34% (m = 100).
- For the right-leaning model (bottom), the right video % does not decrease much (44.73% → 42.40%).
- Overall, more ideologically diverse and balanced content is recommended post-intervention for both models, even though the changes are less pronounced for right-leaning users.

Data Statistics

	Training	Testing	Total
Videos watched	9,930,110	5,393,820	15,323,930
Unique video	23,735	381,153	399,935
Channels	1,256	119,811	120,073