An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes

Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Maria Bielikova

Stuart Heeb
Have you ever noticed...
Let’s take a trip down the YouTube rabbit hole
Andrew Tate First Words After Release Gets REJECTED

183.667 Aufrufe vor 22 Stunden
Andrew Tate First Words After Release Gets REJECTED

► Click Here To Subscribe ► http://goo.gl/Q4Jivn: Mehr ansehen

1.799 Kommentare  Sortieren nach
Recommendations

“Our recommendation system is built on the simple principle of helping people find the videos they want to watch and that will give them value”

Exploration vs. exploitation

https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/
Filter bubbles

Feedback loops in recommendation systems can give rise to “echo chambers” and “filter bubbles” which can narrow a user’s content exposure, and ultimately shift their world view.

5:06 PM · Mar 1, 2019

https://twitter.com/DeepMind/status/1101514121563041792?s=20

generated by DALL-E
Is bubble bursting possible?

Motivation

• Need for independent oversight of personalization behavior
Reference study

• Hussein et al. (2020): *Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube* [1], experiment conducted in mid 2019

• This study’s experiment was conducted in March 2021
Reference study

• Filter bubbles are easily and quickly created

• “YouTube still has a long way to go to mitigate misinformation on its platform” [1]
Audits

Crowdsourcing

- using real user data
- uncontrolled environment
- hard to make comparisons

Sockpuppeting
Audits

Crowdsourcing
- using real user data
- uncontrolled environment
- hard to make comparisons

Sockpuppeting
- using non-human bots
- selection of appropriate seed data
Agents

N. Virginia

06.06.1990

„rather not say“
Agents

• Watches videos for $\leq 30$ mins

• Does not
  • Like
  • Subscribe
  • Comment
  • Act human!
Human factors

• Selective exposure
• Confirmation bias
• Dunning-Kruger effect
Human factors

- Selective exposure
- Confirmation bias
- Dunning-Kruger effect
Human factors

- Selective exposure
- Confirmation bias
- Dunning-Kruger effect
Human factors

• Selective exposure

• Confirmation bias

• Dunning-Kruger effect

“The less you know, the more confident you are”
Experiment

Agent Initialization

Promoting

Debunking

Tear-down
Experiment

- Most popular promoting/debunking videos (seed data)
- Search queries (e.g. “9/11 conspiracy”)
- Wait time between each query
Experiment

• Create the filter bubble
Experiment

• Burst the filter bubble
Experiment

• Clear YouTube history
Topics

- Anti-vaccination
- 9/11
- Chem-trails
- Moon landing
- Flat earth
Metrics

- Score $x_i$ of a single video

\[ x_i = 1 \iff \text{promoting} \]
\[ x_i = 0 \iff \text{neutral} \]
\[ x_i = -1 \iff \text{debunking} \]
Metrics

- **Normalized Score**

\[
NS = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

For recommendations
\[ NS = \frac{1}{n} \sum_{i=1}^{n} x_i \]
\[ = \frac{1}{n} \left( x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} \right) \]
\[ = \frac{1}{10} (1 + 0 - 1 + 1 + 0 + 1 + 0 + 1 - 1 + 1) = 0.3 \]
NS

debunking: $-1$

neutral: $0.3$

promoting: $+1$

\[ NS = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ = \frac{1}{n} (x_1 + x_2 + x_3 + x_4 + x_5) \]

\[ = \frac{1}{10} (1 + 0 - 1 + 1 + 0 + 1 + 0 + 1 - 1 + 1) = 0.3 \]
Metrics

• Search Result Page Misinformation Score

\[
\text{SERP-MS} = \frac{1}{n \cdot (n+1)} \sum_{i=1}^{n} x_i \cdot (n - r_i + 1)
\]

For search results
SERP-MS = \frac{1}{n \cdot (n+1)} \sum_{i=1}^{n} x_i \cdot (n - r_i + 1)

= \frac{1}{n \cdot (n+1)} \left( x_1 \cdot (n - r_1 + 1) + x_2 \cdot (n - r_2 + 1) + x_3 \cdot (n - r_3 + 1) + x_4 \cdot (n - r_4 + 1) + x_5 \cdot (n - r_5 + 1) \right)
SERP-MS

\[
\text{SERP-MS} = \frac{1}{n \cdot (n+1)} \sum_{i=1}^{n} x_i \cdot (n - r_i + 1)
\]

\[
= \frac{1}{15} \left( 1 \cdot (5 - 1 + 1) - 1 \cdot (5 - 2 + 1) + 0 \cdot (5 - 3 + 1) - 1 \cdot (5 - 4 + 1) - 1 \cdot (5 - 5 + 1) \right)
\]

\[
= \frac{1}{15} (5 - 4 + 0 - 2 - 1) = -0.133
\]
SERP-MS

\[
\text{SERP-MS} = \frac{1}{n(n+1)} \sum_{i=1}^{n} x_i \cdot (n - r_i + 1)
\]

\[
= \frac{1}{15} \left( -1 \cdot (5 - 1 - 1) + 0 \cdot (5 - 3 - 1) - 1 \cdot (5 - 4 - 1) - 1 \cdot (5 - 5 + 1) \right)
\]

\[
= \frac{1}{15} (5 - 4 + 0 - 2 - 1) = -0.133
\]
Hypotheses

“How has YouTube’s personalization behavior changed with regards to misinformation videos since the reference study?” [2]
Results

• Comparison with reference study (expecting **better**)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Search results score</th>
<th>Recommendation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/11</td>
<td>n.s.d.</td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Chemtrails</td>
<td>n.s.d.</td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Flat earth</td>
<td>n.s.d.</td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Moon landing</td>
<td>n.s.d.</td>
<td><strong>better</strong></td>
</tr>
<tr>
<td>Anti-vaccination</td>
<td><strong>worse</strong></td>
<td><strong>worse</strong></td>
</tr>
</tbody>
</table>

n.s.d. = not statistically significantly different
Hypotheses

“How does the effect of misinformation filter bubbles change, when debunking videos are watched?” [2]
• Bubble creating behavior (expecting **worse**)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Search results score</th>
<th>Recommendation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/11</td>
<td>n.s.d.</td>
<td>worse</td>
</tr>
<tr>
<td>Chemtrails</td>
<td>n.s.d.</td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Flat earth</td>
<td><strong>better</strong></td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Moon landing</td>
<td>n.s.d.</td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Anti-vaccination</td>
<td>n.s.d.</td>
<td><strong>worse</strong></td>
</tr>
</tbody>
</table>

*n.s.d. = not statistically significantly different*
Results

- Bubble bursting behavior (expecting better)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Search results score</th>
<th>Recommendation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/11</td>
<td>n.s.d.</td>
<td>better</td>
</tr>
<tr>
<td>Chemtrails</td>
<td>n.s.d.</td>
<td>better</td>
</tr>
<tr>
<td>Flat earth</td>
<td>n.s.d.</td>
<td>better</td>
</tr>
<tr>
<td>Moon landing</td>
<td>n.s.d.</td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Anti-vaccination</td>
<td>better</td>
<td>better</td>
</tr>
</tbody>
</table>

n.s.d. = not statistically significantly different
Results

• Comparison to baseline (expecting better)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Search results score</th>
<th>Recommendation score</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/11</td>
<td>n.s.d.</td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Chemtrails</td>
<td>better</td>
<td>better</td>
</tr>
<tr>
<td>Flat earth</td>
<td>better</td>
<td>better</td>
</tr>
<tr>
<td>Moon landing</td>
<td>better</td>
<td>n.s.d.</td>
</tr>
<tr>
<td>Anti-vaccination</td>
<td>better</td>
<td>better</td>
</tr>
</tbody>
</table>

n.s.d. = not statistically significantly different
Outlook

• Srba et al. (2023): Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles [3], continuation of this paper
Conclusion

• YouTube seems to have not fulfilled its pledges

• Bubble bursting is possible, but there are differences between topics
Discussion

• What *is* misinformation?

• How much should YouTube intervene in this matter?

• How strongly should recommendations adhere to human tendencies?

• Does YouTube treat misinformation topics differently?

• Study annotation score vs. YouTube’s “internal scoring”
References


Sources

• Downward stair case: DALL-E (http://labs.openai.com), „a downward spiral into a dark dimension, digital art“ (accessed March 14, 2023)


• Screenshot of tweet: https://twitter.com/DeepMind/status/1101514121563041792?s=20 (accessed March 14, 2023)

• Bubble: DALL-E (http://labs.openai.com), „a soap bubble, that is also a portal to a dark dimension, digital art“ (accessed March 13, 2023)


• Gender-neutral icon: https://thenounproject.com/icon/gender-neutral-147092/ (accessed March 13, 2023)

• Magnifying glass: DALL-E (http://labs.openai.com), „magnifying glass which puts only a part of a document in focus“ (accessed March 19, 2023)