Personalisation Algorithms & Extremist Content Online

Dr Alastair Reed @reed_alastair
Joe Whittaker @CTProject_JW
Fabio Votta @favstats
Seán Looney @_Sean_Looney_

26th June 2019
Slides available here:
xrw-and-algorithms.netlify.com

CYTREC
CYBER THREATS RESEARCH CENTRE

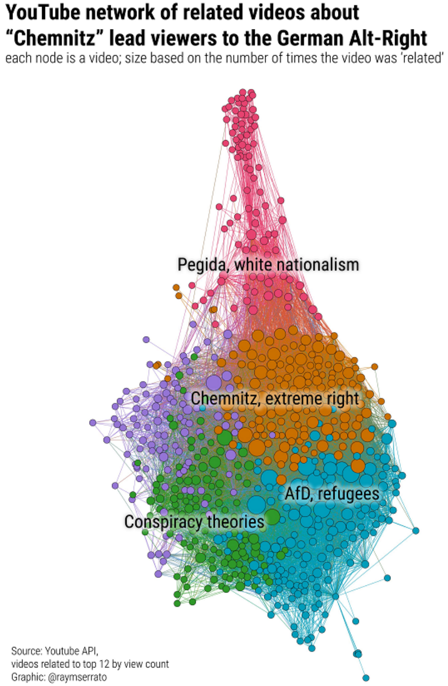
Swansea University
Prifysgol Abertawe

ICCT

What are Filter bubbles?

And what's the problem?

What are Filter bubbles?



YouTube network of related videos about “Chemnitz” lead viewers to the German Alt-Right

each node is a video; size based on the number of times the video was 'related'

Source: Youtube API,
videos related to top 12 by view count
Graphic: @raymserrato

MOSTAFA M. EL-BERMAWY BUSINESS 11.18.16 05:45 AM

YOUR FILTER BUBBLE IS DESTROYING DEMOCRACY

WIRED

THE INTERPRETER

The New York Times

How Everyday Social Media Users Become Real-World Extremists

By Max Fisher and Amanda Taub

April 25, 2018

Everybody's in a Bubble, and That's a Problem

In politics as well as business, people are shaped by who they see—and who they don't. DEREK THOMPSON JAN 25, 2017

The Atlantic

Research so far

and linking filter bubbles to extremism

Research so far

The empirical evidence of a "filter bubble" effect is less clear and decidedly less pessimistic.

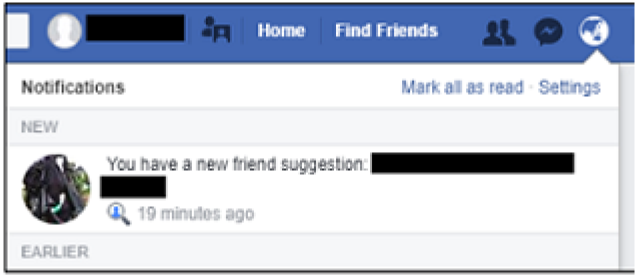
- Study on Facebook suggests filter bubbles are generated less by algorithms than by individual user decisions (Bakshy, Messing and Adamic, 2015).
- Research analysing Google news recommendation suggests essential information is not omitted (Haim, Graefe and Brosius, 2018)
- Personalised recommendations show no reduction in diversity over human editors (Möller et al., 2018).
- Research on Google search results also finds factors such as time of search were more explanatory than prior behaviour and preferences (Courtois, Slechten and Coenen, 2018).

Why the discrepancy?

- "Echo chamber about echo chambers" (Guess et al. 2018)

Filter Bubbles and Extremism

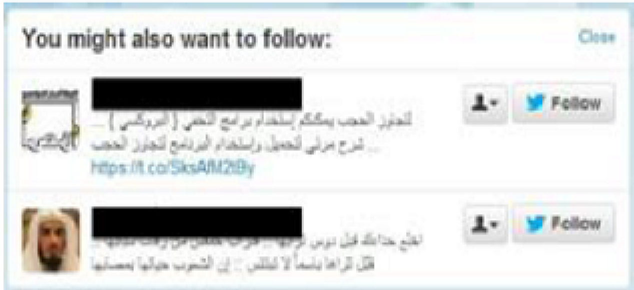
There is a paucity of research studying the effects of personalisation algorithms on extremist content.



You have a new friend suggestion: [redacted]

19 minutes ago

Figure 6: Facebook sending a notification for a new suggested friend - an active IS fighter in the Philippines



[redacted]

لتجاوز الحجب يمكنك إستخدام برنامج التخفي (البروكسي) ... شرح مرئي لتحميل وإستخدام البرنامج لتجاوز الحجب ...

https://t.co/SksAfM2t9y

[redacted]

[illegible]

- YouTube's recommended videos can propel users into an immersive bubble of right-wing extremism (O’Callaghan et al., 2015)
- Twitter’s “Who to Follow” suggested violent extremist Islamist groups if the user followed al-Qaeda affiliated group (Berger, 2013)
- Facebook’s “Recommended Friends” function had likely actively connected at least two Islamic State supporters in SE Asia (Waters and Postings, 2018)

The architecture of the platforms may facilitate closer interactions than would otherwise exist.

Research Question & Design

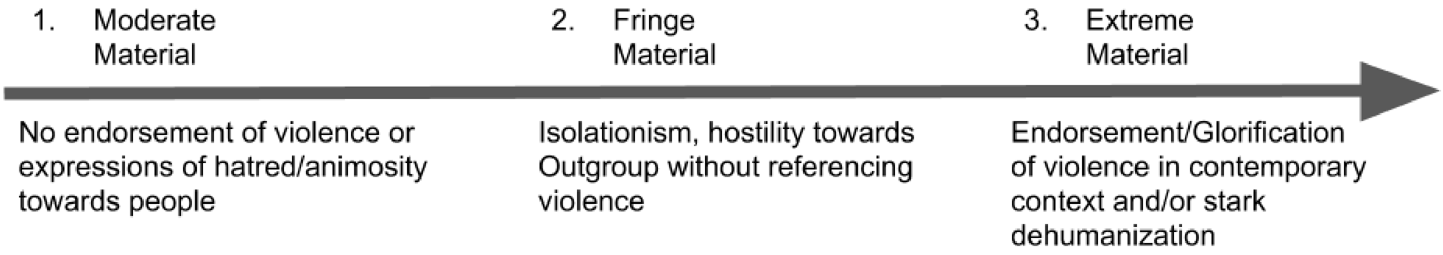
Research Question

> Do algorithms promote extremist material once a user begins to interact with such content?

How to measure extremist content?

- Hand-coding content with Holbrook's Extremist Media Index

Holbrook's Extremism Media Index (2015)



1. Moderate Material

No endorsement of violence or expressions of hatred/animosity towards people

2. Fringe Material

Isolationism, hostility towards Outgroup without referencing violence

3. Extreme Material

Endorsement/Glorification of violence in contemporary context and/or stark dehumanization

Research Design & Data Collection

- YouTube/Reddit Research design
  - Created **THREE** identical accounts
  - All follow same 10 XRW channels/subreddtis; 10 Neutral.[1]
  - Each account interacts with different kinds of content
- Collecting timelines *two times a day* for *two weeks* (28 sessions in total)
  - 2019-01-21 and ended on 2019-02-04

Each account does nothing for a week and after Session 14 we apply different treatments:

1. *Neutral Interaction Account* mostly interacts with neutral content
2. *Extreme Interaction Account* mostly interacts with extreme content
3. *Baseline Account* does nothing to establish a baseline

- On YouTube: Pull 18 Recommended Videos of each account
- On Reddit: 25 threads on the *"Best"* personalized timeline for each account
- Every piece of content gets a unique rank per session and an Extremist Media Index (EMI) Score (Holbrook, 2015)

[1] We follow the policy of Vox Pol and Berger (2018) in not identifying the names of accounts in this research, both for reasons of potentially increasing exposure and privacy.

Research Design & Methods

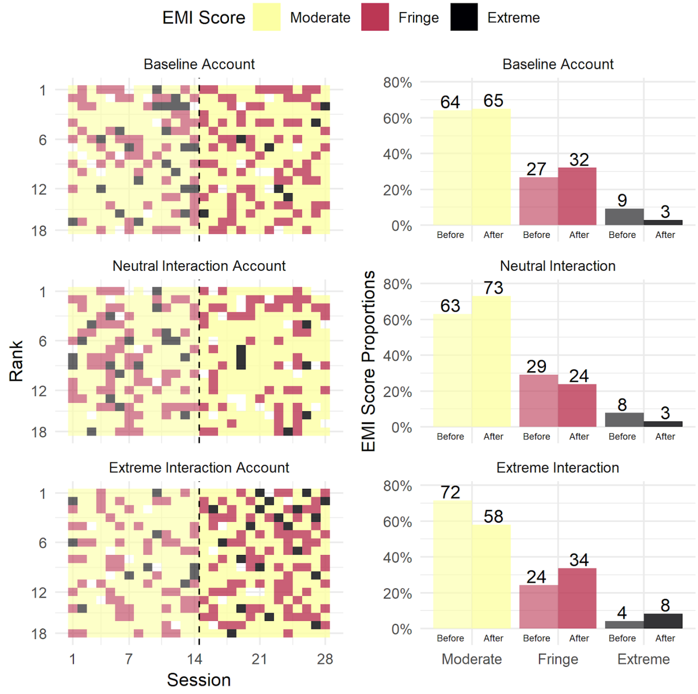
Expected Relationships

- Frequency of extreme material increases after interacting with extreme content
- Extreme content is prioritized by the algorithm when interacting with extreme material

Methods

- Count data modeled with (quasi-)poisson regression
- Non-Parametric t-tests to estimate ranking differences
- Satisfactory interrater reliability between two individuals
    - Krippendorff's alpha: 0.77 across all platforms

Results

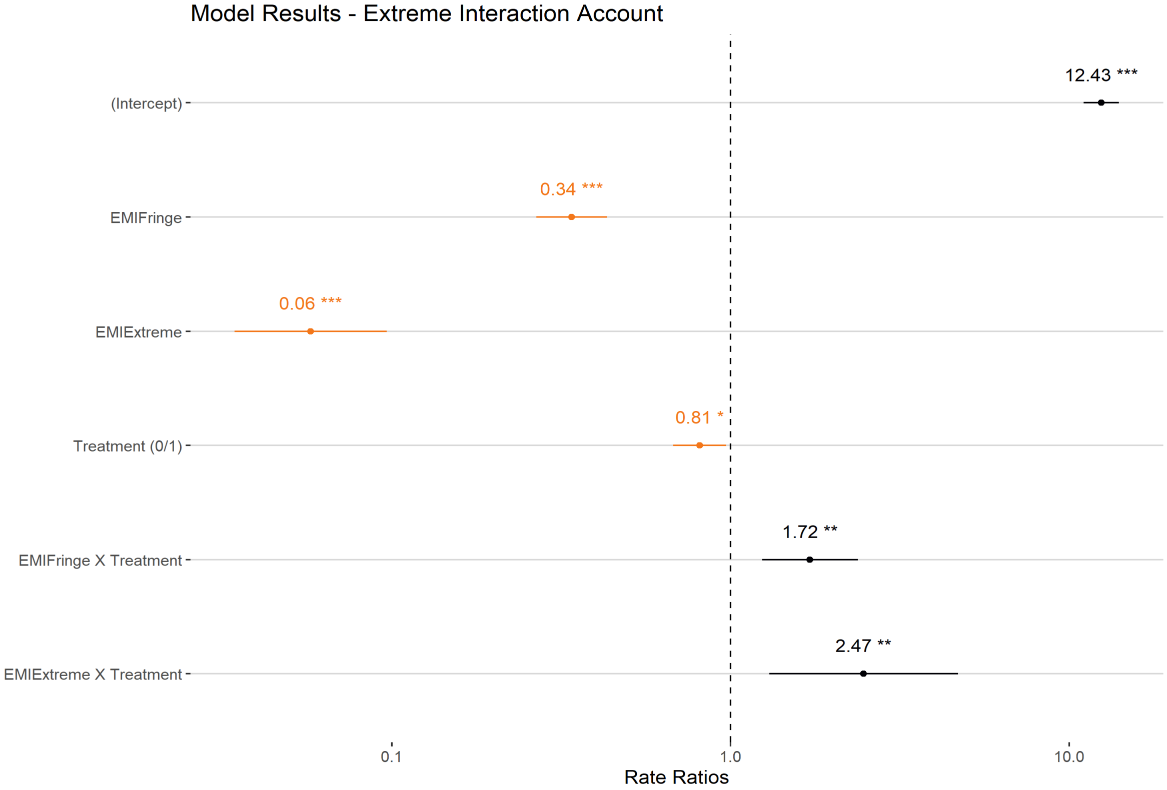


Overview of the data:

Of the 1443 videos coded on YouTube (749 unique)

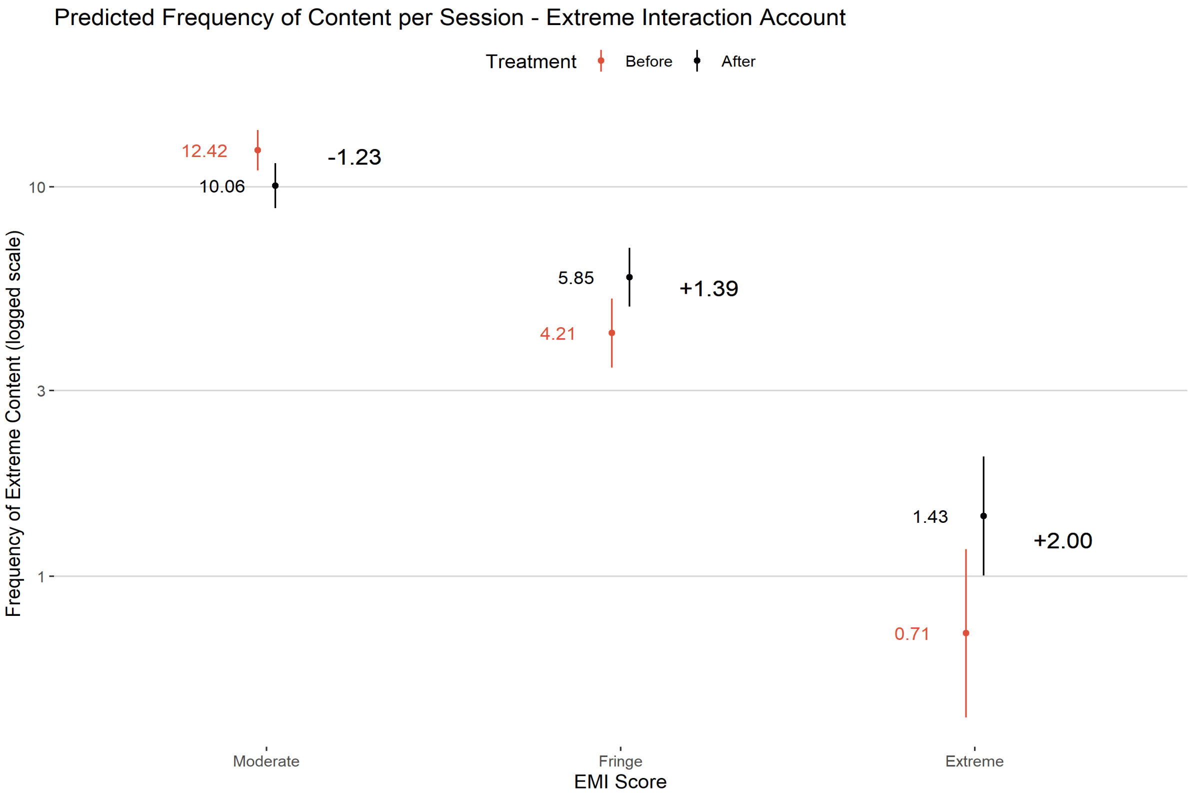
- 65.77% moderate
- 28.34% fringe
- 5.89% extremist

Figure on the left shows the EMI scores for each session with a rank from one to eighteen, depending on where the video appears on the “Recommended Videos” section. Figure on the right shows the percentage distribution of the three categories of content before and after each treatment.



After **extreme interaction** treatment:

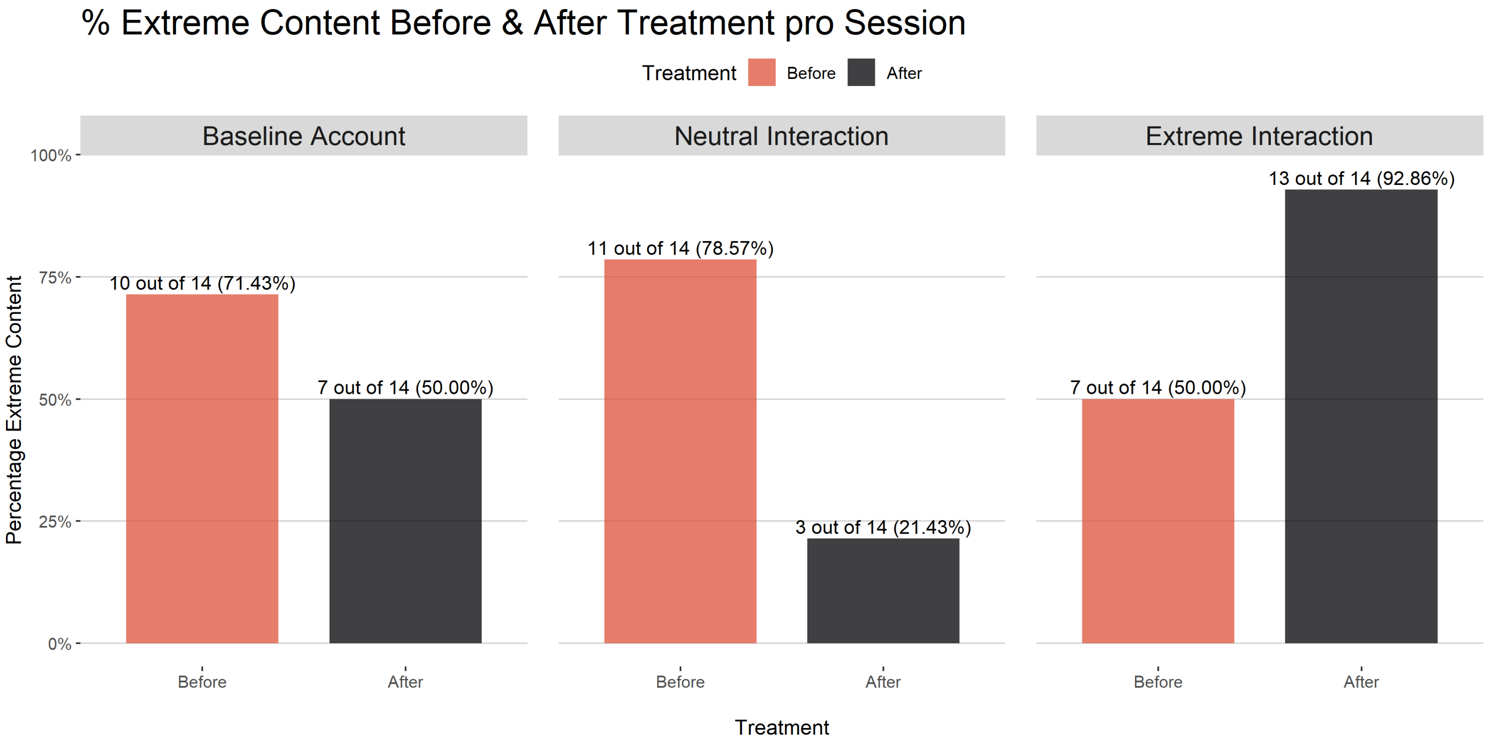
- The incident rate for *fringe content* is 1.72 times the incident rate for the reference group
- The incident rate for *extreme content* is 2.47 times the incident rate for the reference group



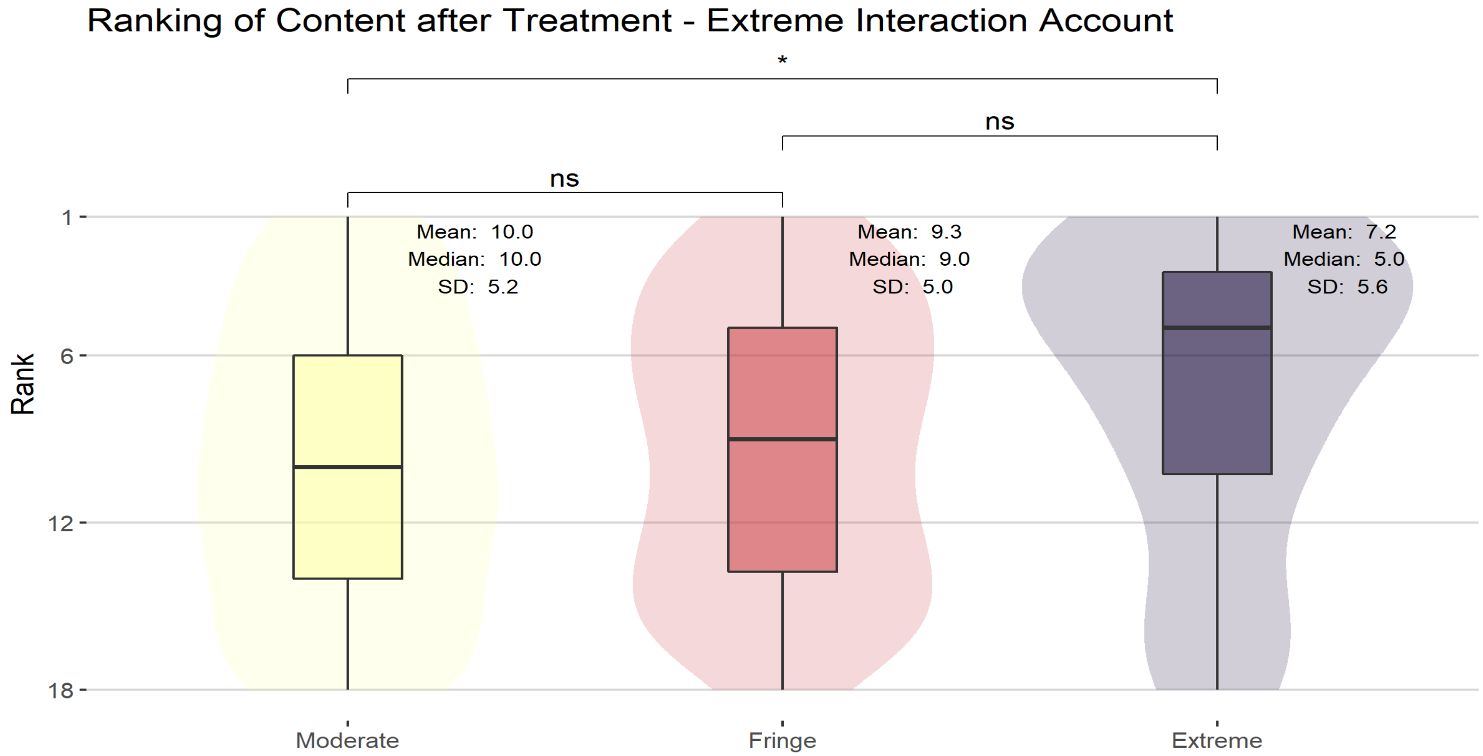
After **extreme interaction** treatment:

- Fringe content 1.37 (p < 0.01) times more likely than before
- Extreme content 2.00 (p < 0.01) times more likely than before

You Tube

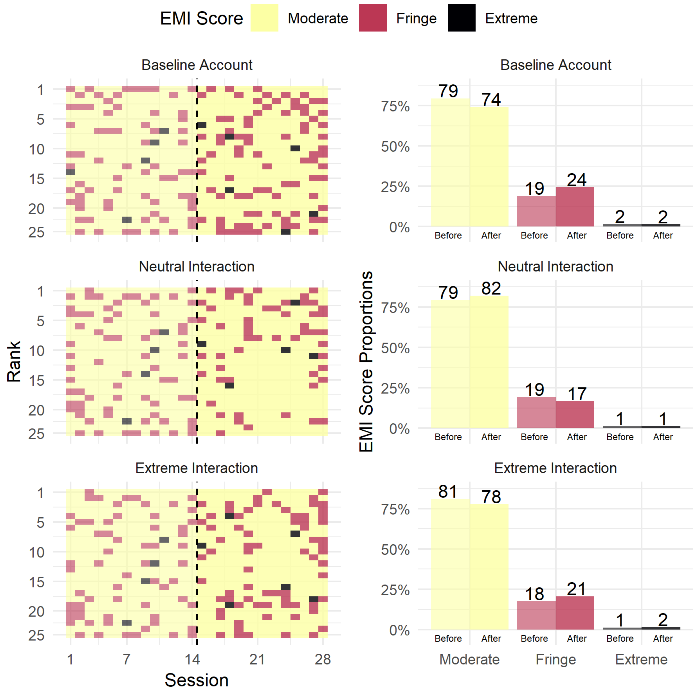


- In the **neutral interaction account**
  - **only three sessions** had an extreme content piece after interaction
- In the **extreme interaction account**
  - **all but one session** had an extreme content piece after interaction



- In the **extreme interaction account**
  - Extreme content ranked sig. higher (p = 0.028) than moderate
  - Almost all extreme content shows up in the upper half (< 8) of the recommendations (Median Rank = 5)

reddit

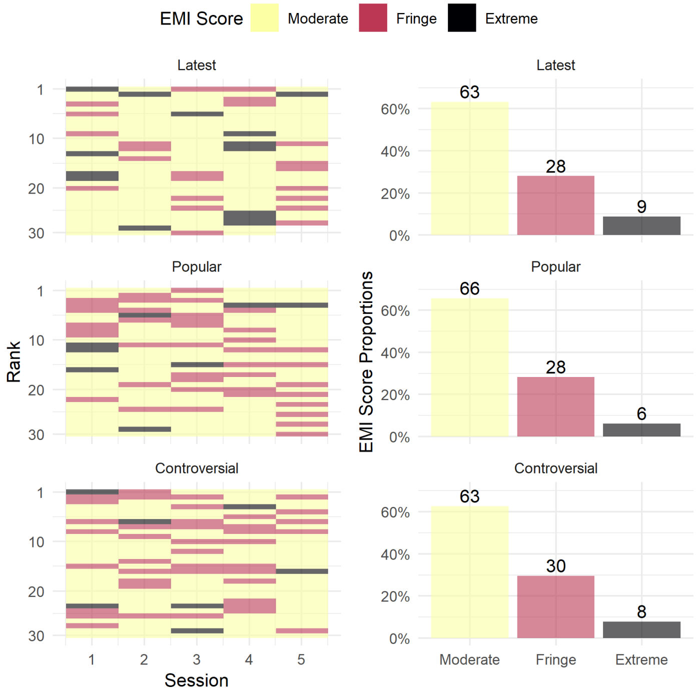


Of the 2100 posts coded on Reddit (834 unique)

- 78.76% Moderate
- 19.81% Fringe
- 1.43% Extreme

No sig. observable in-/decrease of content

No sig. difference in ranking of content



Different setup due to technical difficulties

3 Different News Feeds: “Popular”, “Controversial”, and “Latest”

3 Topics: “Politics”, “News”, “Humour”

Collected data over five sessions

1271 Rated posts (746 unique)

- 63.73% Moderate
- 28.8% Fringe
- 7.47% Extreme

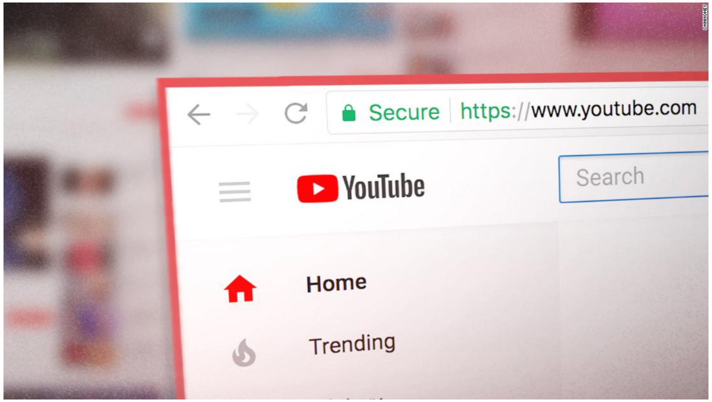
No statistically sig. difference between extreme & non-extreme content in any of the timelines

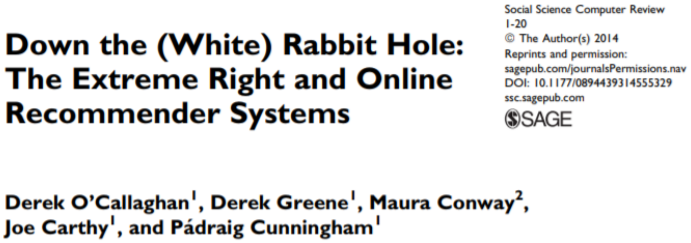
Discussion

Discussion

> Do algorithms promote extremist material once a user begins to interact with such content?

- Only YouTube has an effect which, after engaging with extreme content, prioritises it further.
- Users which engage with extreme and fringe content are more likely to be recommended more of the same.
- Extreme content is pushed up the ranking of recommended videos on YouTube
- Support for O'Callaghan et al. (2015)





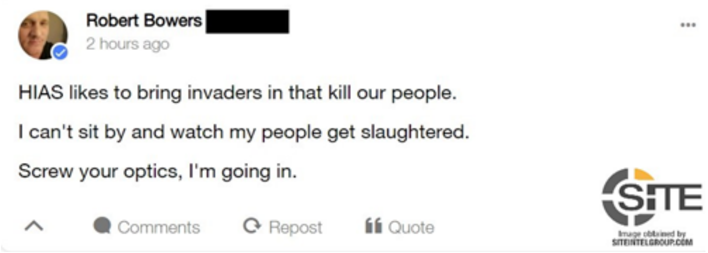
Social Science Computer Review
1-20
© The Author(s) 2014
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0894439314555329
ssc.sagepub.com
SAGE

Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems

Derek O'Callaghan[1], Derek Greene[1], Maura Conway[2], Joe Carthy[1], and Pádraig Cunningham[1]

Discussion

- Safe haven for right wing extremists and home to terrorists such as Pittsburgh Synagogue Shooter
- Anecdotally, by far the most extreme of the platforms.
- Lack of evidence of Gab's algorithm suggests that it is the user's choices which are responsible for this environment.



Robert Bowers [redacted]
2 hours ago

HIAS likes to bring invaders in that kill our people.

I can't sit by and watch my people get slaughtered.

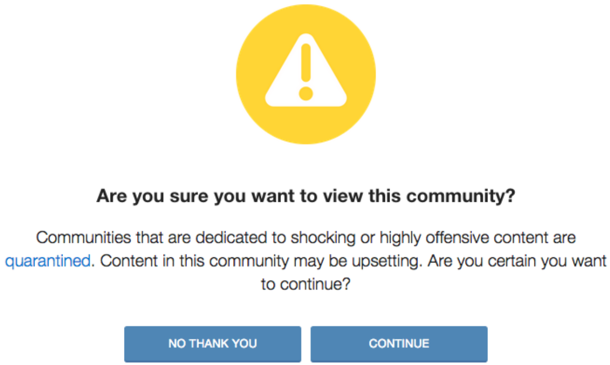
Screw your optics, I'm going in.

Recommendations

Recommendations

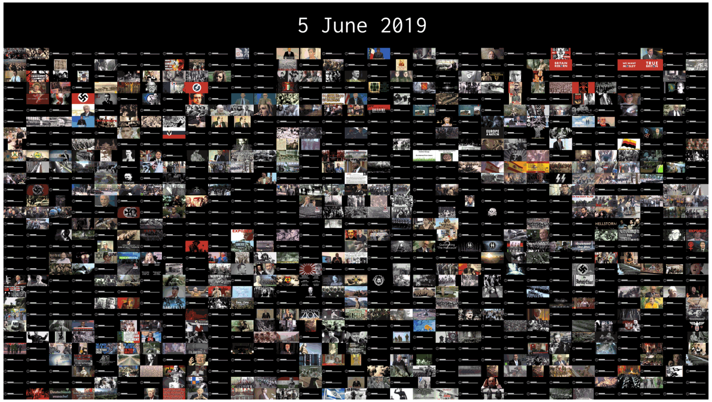
Removing Problematic Content from Recommendations

- Content which does not clearly violate site rules or policies
- Google's "limited features" policy
- Reddit's "Quarantine" System
- Opt in content
- No monetisation or recommendation
- Constructive balance between freedom of speech and harmful content.
- Need for Clarity and Consistency



Are you sure you want to view this community?

Communities that are dedicated to shocking or highly offensive content are quarantined. Content in this community may be upsetting. Are you certain you want to continue?



OILAB: Availability of YouTube videos posted in Nazi threads on 4chan/pol/ on 5 and 6 June 2019.

Recommendations

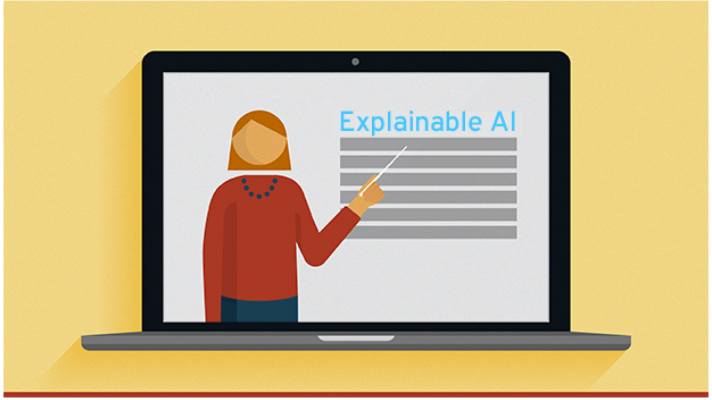
Ensuring Video Recommendations are from Quality Sources

- Provide users with more context and alternative perspectives
- Google introduced changes to make quality count and give more context to searches.
- Provide high quality sources on the same topic
- Jigsaw’s redirect method as a model.

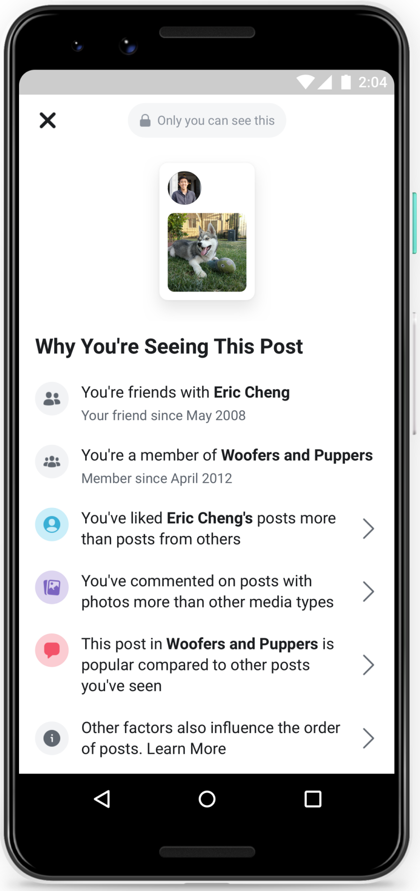
Recommendations

Greater Transparency

- Users should have a clear option to request why content has been recommended to them.
- Opportunity for Explainable AI
- Facebook's "Why am I seeing this ad/post"



Greater Transparency



2:04

Only you can see this

Why You're Seeing This Post

You're friends with **Eric Cheng**
Your friend since May 2008

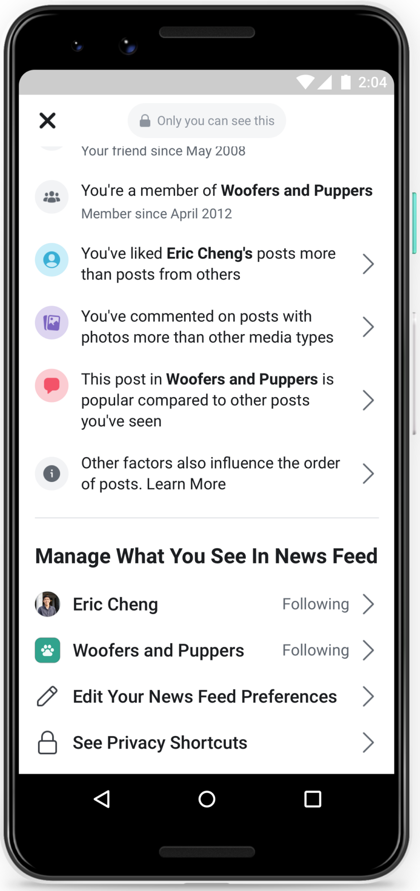
You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with photos more than other media types

This post in **Woofers and Puppers** is popular compared to other posts you've seen

Other factors also influence the order of posts. Learn More



2:04

Only you can see this

Your friend since May 2008

You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with photos more than other media types

This post in **Woofers and Puppers** is popular compared to other posts you've seen

Other factors also influence the order of posts. Learn More

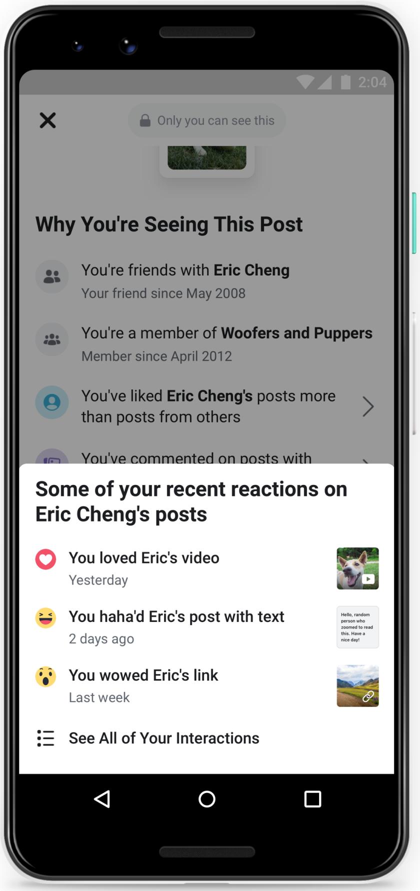
Manage What You See In News Feed

Eric Cheng — Following

Woofers and Puppers — Following

Edit Your News Feed Preferences

See Privacy Shortcuts



2:04

Only you can see this

Why You're Seeing This Post

You're friends with **Eric Cheng**
Your friend since May 2008

You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with

Some of your recent reactions on Eric Cheng's posts

You loved Eric's video
Yesterday

You haha'd Eric's post with text
2 days ago

You wowed Eric's link
Last week

See All of Your Interactions

Source: Facebook Newsroom, 31st March 2019

Future Research

- More accounts over a longer period of time
- Research Project constrained by the number and type of social media platforms that we could research.
- Closed nature of the platform (Facebook)
- Terms of Service Restrictions (Twitter)
- Increasing knowledge gap which can only be answered through close collaboration with social media companies.

Thank you for listening!

Literature

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348(6239), 1130-1132.

Berger, J. M. (2013) ‘Zero Degrees of al Qaeda’, Foreign Policy, (August 14). Available at: http://foreignpolicy.com/2013/08/14/zero-degrees-of-al-qaeda/.

Courtois, C., Slechten, L., & Coenen, L. (2018). Challenging Google Search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. Telematics and Informatics, 35(7), 2006-2015.

Guess, A., Lyons, B., Nyhan, B., & Reifler, J. (2018). Avoiding the echo chamber about echo chambers: Why selective exposure to like-minded political news is less prevalent than you think. Document of the Knight Foundation. Retrieved from:
https://www.researchgate.net/publication/330144926_Avoiding_the_echo_chamber_abo
minded_political_news_is_less_prevalent_than_you_think

Haim, M., Graefe, A., & Brosius, H. B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. Digital Journalism, 6(3), 330-343.

Literature

Holbrook, D. (2015). Designing and Applying an ‘Extremist Media Index’. Perspectives On Terrorism, 9(5). Retrieved from http://www.terrorismanalysts.com/pt/index.php/pot/article/view/461

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. Information, Communication & Society, 21(7), 959-977.

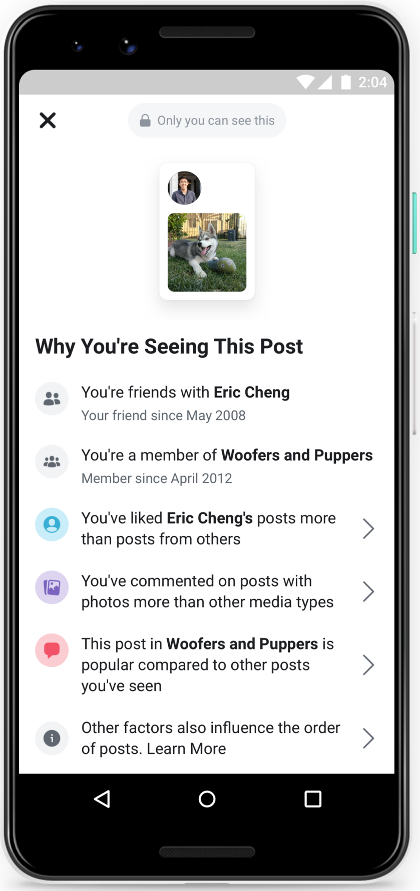
O’Callaghan, Derek, et al. "Down the (white) rabbit hole: The extreme right and online recommender systems." Social Science Computer Review 33.4 (2015): 459-478.

Vīķe-Freiberga, V., Däubler-Gmelin, H., Hammersley, B., Pessoa Maduro, L.M.P. (2013). A free and pluralistic media to sustain European democracy. Retrieved from http://ec.europa.eu/digital-agenda/sites/digital-agenda/files/HLG%20Final%20Report.pdf

Waters, G. and Postings, R. (2018) Spiders of the Caliphate: Mapping the Islamic State’s Global Support Network on Facebook, Counter-Extremism Project.

Appendix

Facebook adds "Why am I seeing this" to posts



2:04

✕ Only you can see this

Why You're Seeing This Post

You're friends with **Eric Cheng**
Your friend since May 2008

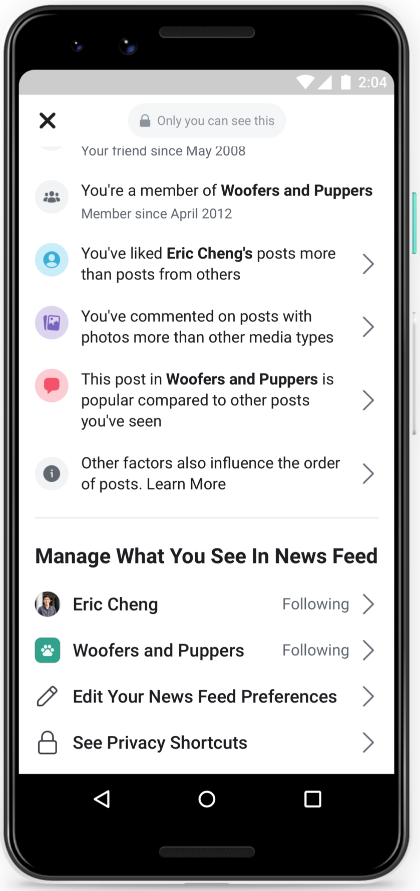
You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with photos more than other media types

This post in **Woofers and Puppers** is popular compared to other posts you've seen

Other factors also influence the order of posts. Learn More



2:04

✕ Only you can see this

Your friend since May 2008

You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with photos more than other media types

This post in **Woofers and Puppers** is popular compared to other posts you've seen

Other factors also influence the order of posts. Learn More

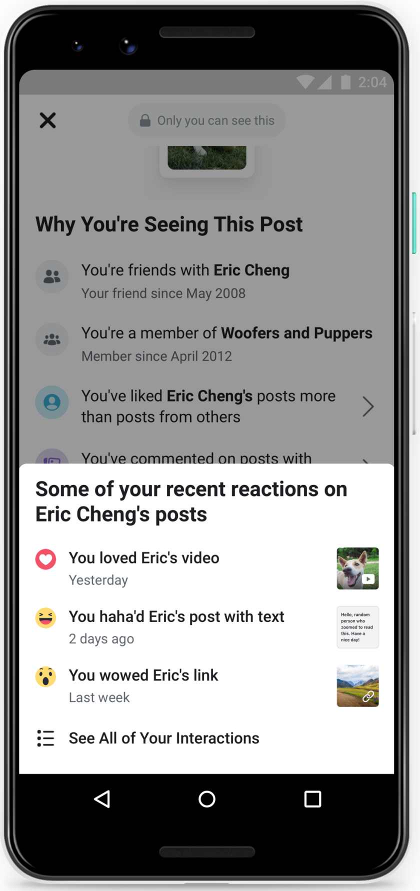
Manage What You See In News Feed

Eric Cheng — Following

Woofers and Puppers — Following

Edit Your News Feed Preferences

See Privacy Shortcuts



2:04

✕ Only you can see this

Why You're Seeing This Post

You're friends with **Eric Cheng**
Your friend since May 2008

You're a member of **Woofers and Puppers**
Member since April 2012

You've liked **Eric Cheng's** posts more than posts from others

You've commented on posts with

Some of your recent reactions on Eric Cheng's posts

You loved Eric's video
Yesterday

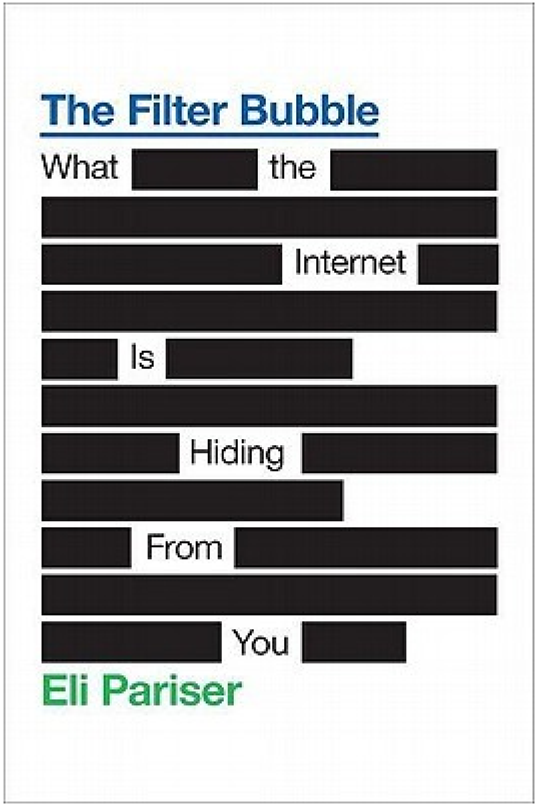
You haha'd Eric's post with text
2 days ago

You wowed Eric's link
Last week

See All of Your Interactions

Source: Facebook Newsroom, 31st March 2019

What are Personalisation Algorithms?


The Filter Bubble

What the Internet Is Hiding From You

Eli Pariser

Pariser 2011

- Algorithms are responsible for content that users see in their feeds
- Eli Pariser suggests that they can create a “filter bubble” effect or “autopropaganda”
- by controlling what users do and do not see it can – and is in fact, designed to – dramatically amplify confirmation bias
- The pre-filtering of content leads to bubbles in which people never view or read about opposing viewpoints
- Creates seperated spaces that make communication between opposing viewpoints harder: undermines democracy itself

What are Personalisation Algorithms?

The EU Group on Media Freedom and Pluralism notes that:

> Increasing filtering mechanisms makes it more likely for people to only get news on subjects they are interested in, and with the perspective they identify with. It will also tend to create more insulated communities as isolated subsets within the overall public sphere. […] Such developments undoubtedly have a potentially negative impact on democracy.

Vīķe-Freiberga, Däubler-Gmelin, Hammersley, & Pessoa Maduro, 2013, p. 27

Filter bubbles are considered a concern at the highest level.

How to measure extremist content?

Multiple pathways:

- Sentiment analysis can be used to identify extremist authors (Scrivens et al., 2018)
- Topic models to identify (far-right) extremist content (O’Callaghan et al., 2015)
- Hand-coding content for example with Holbrook's Extremist Media Index (Holbrook, 2015)
- Manually labelled data can also be used as training dataset for machine learning models

YouTube

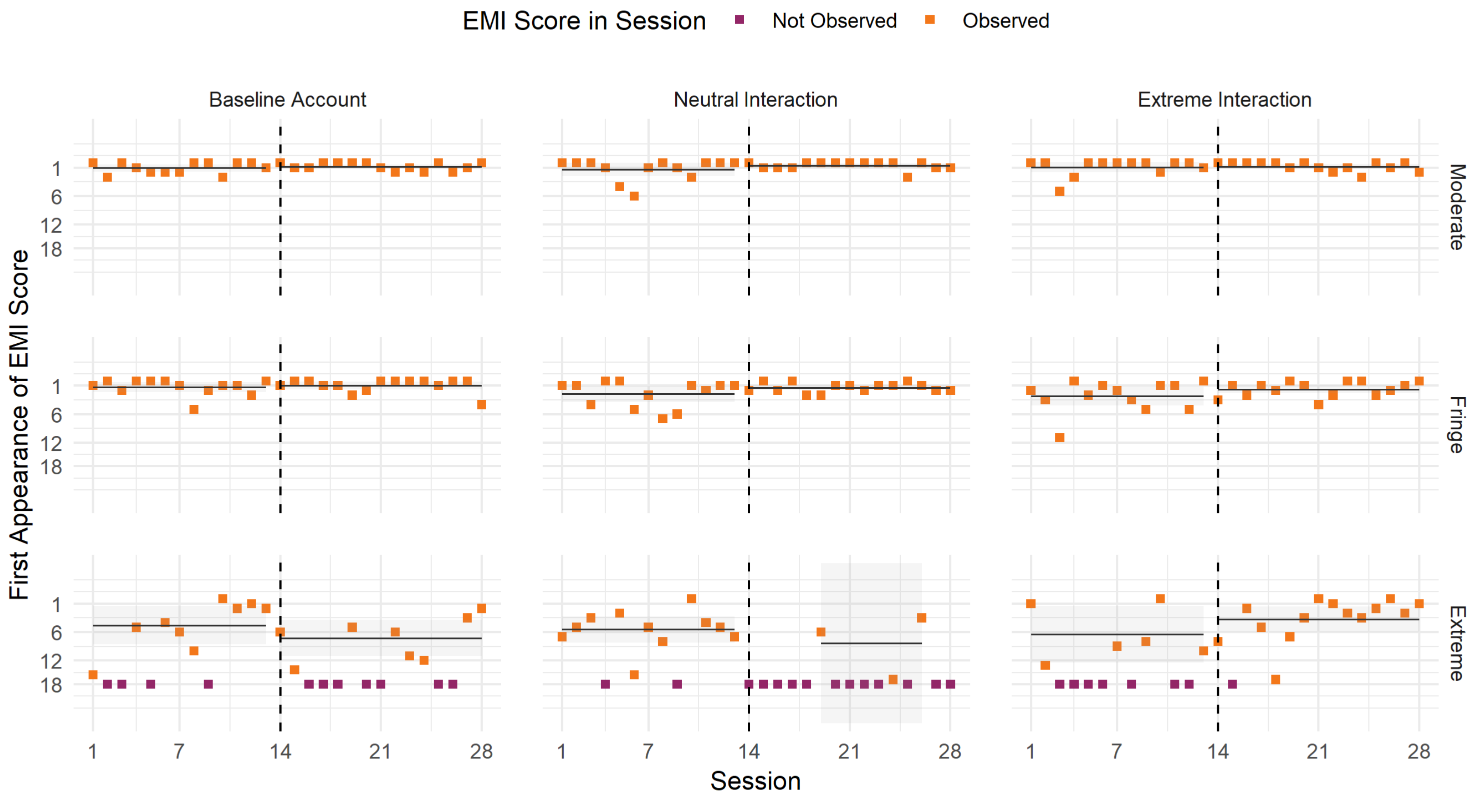


Figure shows the first appearance of a piece of content (Moderate, Fringe or Extreme) for each session.

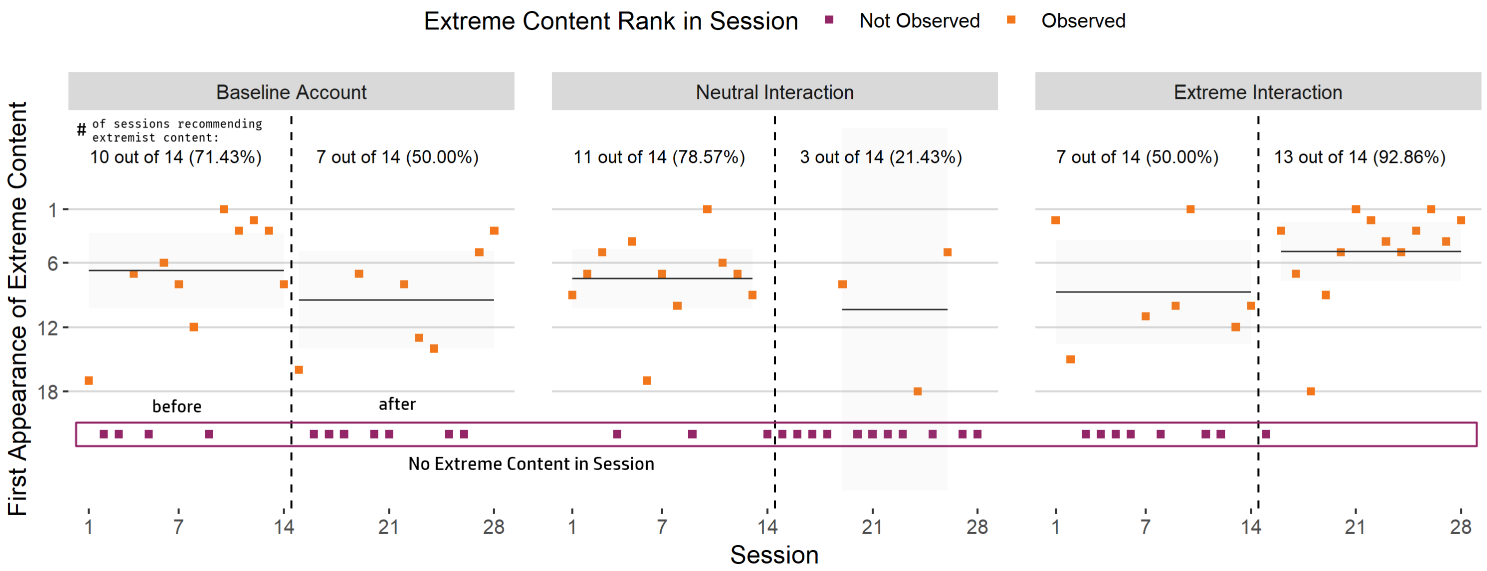


Figure shows the first appearance of an **extreme** content piece for each session.

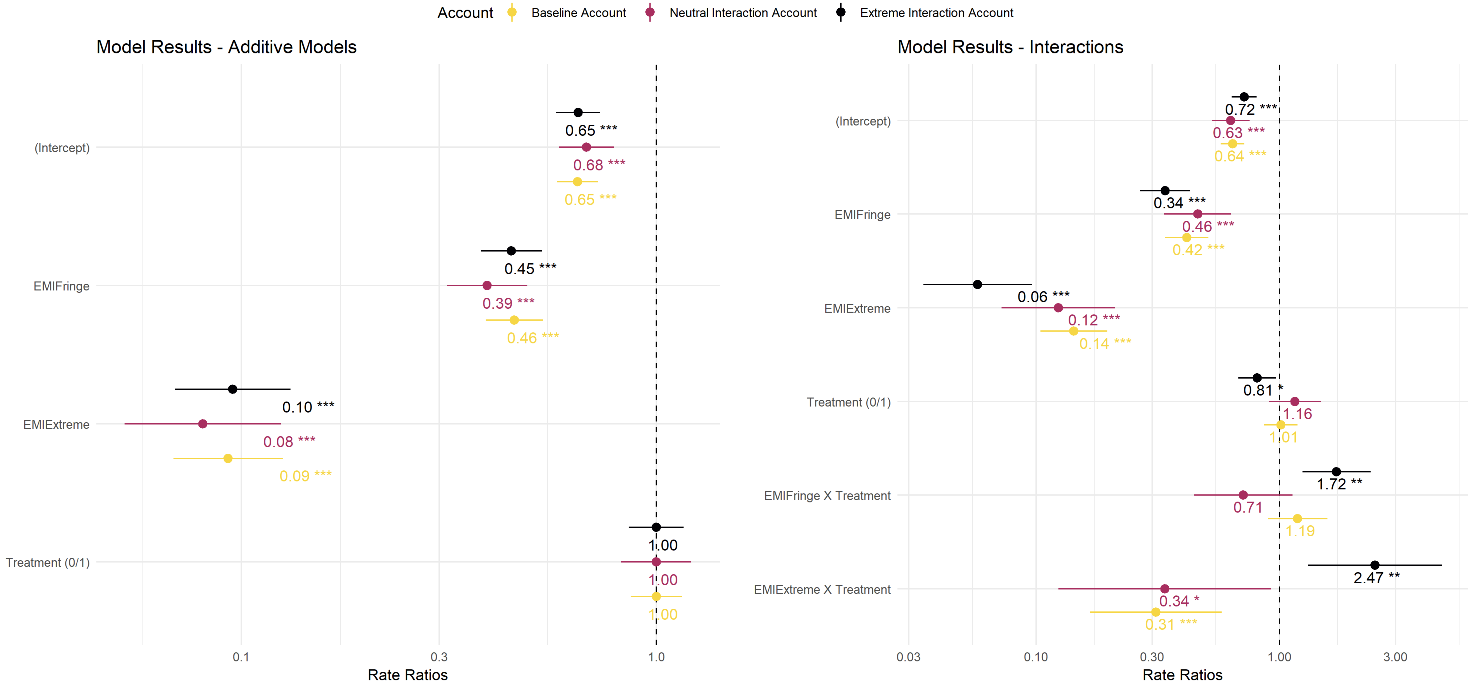
- In the *neutral interaction account*
    - **only three sessions** had an extreme content piece after interaction

- In the *extreme interaction account*

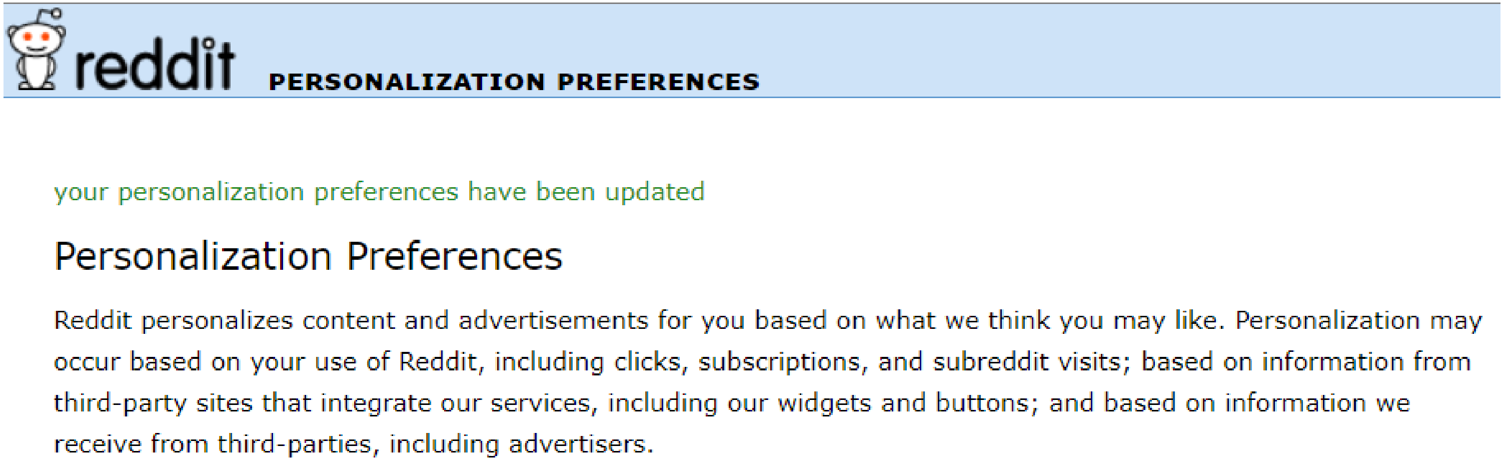
    - **all but one session** had an extreme content piece after interaction
    - Almost all content shows up in the upper half (< 8) of the recommendation list

(Median Rank = 5)

Poisson Regression Results:



Reddit Personalisation



PERSONALIZATION PREFERENCES

your personalization preferences have been updated

Personalization Preferences

Reddit personalizes content and advertisements for you based on what we think you may like. Personalization may occur based on your use of Reddit, including clicks, subscriptions, and subreddit visits; based on information from third-party sites that integrate our services, including our widgets and buttons; and based on information we receive from third-parties, including advertisers.