Royal United Services Institute
for Defence and Security Studies

ICCT International Centre for Counter-Terrorism - The Hague

Swansea University Prifysgol Abertawe

Global Research Network on Terrorism and Technology: Paper No. 8

Radical Filter Bubbles

Social Media Personalisation Algorithms and Extremist Content

Alastair Reed, Joe Whittaker, Fabio Votta and Seán Looney

Key Findings

- Investigations were conducted on three social media platforms – YouTube, Reddit and Gab – to establish whether the sites' recommender systems promote extremist material.
- The recommender system of one social media platform – YouTube – prioritises extreme right-wing material after interaction with similar content, supporting the findings of a previous study.
- The authors do not find evidence of this effect on either Reddit or Gab. This is significant as Gab has been identified by many as a safe haven for extreme right-wing movements. This finding suggests that it is the users, rather than the site architecture, which drives extremist content on Gab.

Recommendations

- Social media platforms should consider removing problematic content that does not break their terms of service from the recommender system, effectively quarantining it.
- Platforms should prioritise recommending high-quality sources (such as news outlets with established editorial departments) when users are interacting with problematic or controversial content that does not clearly violate their policies.
- Social media platforms should provide greater transparency in social media recommendations, including a clearly identifiable 'why am I being recommended this?' option on platforms.
- Further research should be conducted, particularly on closed platforms or sites that currently do not permit such research in their terms of service. This can only be done in close collaboration with social media platforms.

Introduction

Social media platforms and the algorithms that drive them are now an inescapable part of contemporary life. They are the hallmark of the Web 2.0 era, differentiating dynamic, smart platforms from the static websites and fora of the internet's past.[1] These algorithms are responsible for the content users see in their feeds, the content they are recommended and the products they see advertised. Despite this, little is known about the operational workings of these algorithms or their effects on users, precisely because they are central

1. Taina Bucher, 'Want to be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook', *New Media & Society* (Vol. 14, No. 7, 2012), pp. 1164–80.

to platforms' business models.[2] This is even more the case with regards to extremist content and algorithms' roles in trajectories towards terrorism. This lack of understanding is problematic as users who are engaging with extreme content online could, potentially, be recommended further extreme content that they would not have otherwise seen, while dissenting viewpoints could be harder to reach.

The aim of this paper is to assess whether social media platforms' personalisation algorithms promote extremist material. The paper offers an exploratory analysis of recommender systems (for example, YouTube's Recommended tab on its homepage, which offers video suggestions based on a number of different variables) of three social media platforms – YouTube, Reddit and Gab – when put into contact with far-right extremist content. The paper finds evidence that only one platform – YouTube – prioritises extremist material by the recommender system.

The Filter Bubble

Despite the workings of social media personalisation algorithms being unclear to the general public, many have expressed warnings that they may have a malign effect. Eli Pariser, who coined the term 'filter bubble', described them as 'autopropaganda' and argued that they control what users do and do not see. He also argues that they can – and are in fact designed to – dramatically amplify confirmation bias.[3] Others have warned that their human creators may pass on biases,[4] and that platforms have conflated the distinction between user satisfaction and retention,[5] which may cause important events to be filtered out if they do not fit the model of user-retention.[6] The EU Group on Media Freedom and Pluralism has warned that '[i]ncreasing 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 ... Such developments undoubtedly have a potentially negative impact on democracy'.[7]

2. Michael A DeVito, 'From Editors to Algorithms', *Digital Journalism* (Vol. 5, No. 6, 2016), pp. 1–21.
3. Eli Pariser, *The Filter Bubble: What the Internet Is Hiding from You* (London: Penguin, 2011).
4. Engin Bozdag, 'Bias in Algorithmic Filtering and Personalization', *Ethics and Information Technology* (Vol. 15, No. 3, September 2013), pp. 209–27.
5. Nick Seaver, 'Captivating Algorithms: Recommender Systems as Traps', *Journal of Material Culture* (29 December 2018), doi: 10.1177/1359183518820366.
6. Philip M Napoli, 'Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers', *Telecommunications Policy* (Vol. 39, No. 9, October 2015), pp. 751–60.
7. Vaira Vīke-Freiberga et al., 'A Free and Pluralistic Media to Sustain European Democracy', Report of the High Level Group on Media Freedom and Pluralism,

Similarly, the recent UK Government Online Harms White Paper expresses concern that they cause users to only be presented with one type of information, and restricts access to cross-cutting viewpoints.[8] Clearly, there is concern, including at the highest policy level, that personalisation algorithms may have a problematic filter bubble effect.

Empirical evidence of a filter bubble effect, however, is far less clear. Findings from a study on Facebook suggest that although personalisation effects do occur and have a filtering effect, it is smaller than users' own personal choices.[9] Similarly, an experimental study found that customisability on news sites did increase political polarisation, but when both user-driven and system-driven customisability were present, users' choices mitigated the effect of the latter.[10] Other research analysing news recommendation plays down the importance of the filter bubble effect, suggesting that it does not filter out essential information and that personalised recommendations show no reduction in diversity over human editors.[11] Finally, research on Google search results also downplays such an effect and find that factors such as time of search were more explanatory than prior behaviour and preferences.[12]

Filter Bubbles and Extremism

There is a paucity of research studying the effects of personalisation algorithms on extremist content. The most relevant for this study is an investigation into extreme right-wing videos on YouTube, which finds that users who follow the recommender system can be propelled into an immersive ideological bubble within a few clicks.[13] In 2013, it was found that for users following Al-Qa'ida

January 2013, p. 27, <https://ec.europa.eu/digital-single-market/sites/digital-agenda/files/HLG%20Final%20Report.pdf>, accessed 8 July 2019.

8. HM Government, *Online Harms White Paper* (London: The Stationary Office, 2019).
9. Eytan Bakshy, Solomon Messing and Lada Adamic, 'Exposure to Ideologically Diverse News and Opinion on Facebook', *Science Express*, (2015), pp. 1–5.
10. Ivan Dylko et al., 'Impact of Customizability Technology on Political Polarization', *Journal of Information Technology and Politics*, (Vol. 15, No. 1, 2018), pp. 19–33.
11. Judith Möller et al., 'Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and Their Impact on Content Diversity', *Information, Communication & Society*, (Vol. 21, No. 7, 2018), pp. 959–77; Mario Haim, Andreas Graefe and Hans-Bernd Brosius, 'Burst of the Filter Bubble?: Effects of Personalization on the Diversity of Google News', *Digital Journalism* (Vol. 6, No. 3, 2018), pp. 330–43.
12. Cédric Courtois, Laura Slechten and Lennert Coenen, 'Challenging Google Search Filter Bubbles in Social and Political Information: Disconforming Evidence from a Digital Methods Case Study', *Telematics and Informatics* (Vol. 35, No. 7, 2018), pp. 2006–15.
13. Derek O'Callaghan et al., 'Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems', *Social Science Computer Review* (Vol. 33,

affiliate Jabhat Al-Nursa, Twitter's Who to Follow recommendation would suggest a number of other violent extremist accounts.[14] Similarly, research on Southeast Asian supporters of Daesh (also known as the Islamic State of Iraq and Syria, ISIS) found that Facebook's Recommended Friends function had actively connected at least two sympathisers.[15]

Several inferences can be drawn from this research. First, although there is little empirical evidence of a filter bubble effect of extremist content, what does exist suggests it may be increasing the possibility of system-driven promotion of such material. Second, while there is some research analysing the effects of personalisation algorithms and violent extremist content, it mostly looks at the potential content that users could possibly view, rather than the actual behaviours of violent extremists, although the finding that Facebook's Recommended Friends function had actively connected supporters does seek to explain how recommender systems affect behaviour. Third, these three studies are focused on recommender systems, which are optional for the user and do not focus on compulsory aspects such as timelines or news feeds. Finally, the studies were conducted on three of the biggest social media platforms: YouTube; Twitter; and Facebook. While there is good reason to research the largest platforms, there is also value in investigating smaller platforms, especially given the wide array of terrorist and extremist content elsewhere.[16]

Methodology: Data Collection

To assess whether there is a filter bubble effect, three separate investigations are conducted in which the promotion of extreme content by the platforms' recommender systems are tested.

The authors initially identified four social media platforms with different site architectures and different algorithms: Twitter; YouTube; Reddit; and Gab. Unfortunately, it was judged that it would break Twitter's applications programming interface rules to use the interface to measure the performance of the site, which is not permitted,[17] and as such it is decided against including

No. 4, 2015), pp. 459–78.

14. J M Berger, 'Zero Degrees of Al Qaeda', *Foreign Policy*, 14 August 2013.
15. Gregory Waters and Robert Postings, 'Spiders of the Caliphate: Mapping the Islamic State's Global Support Network on Facebook', Counter Extremism Project, 2018, <https://www.counterextremism.com/sites/default/files/Spiders%20of%20the%20Caliphate%20%28May%202018%29.pdf>, accessed 10 July 2019.
16. Maura Conway et al., *Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts* (Dublin: Vox-Pol Network of Excellence, 2017).
17. Twitter, 'Developer Terms: More about Restricted Uses of the Twitter API', <https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.html>, accessed 10 July 2019.

the site in the analysis. The three remaining platforms have different ways in which users are offered data by the personalisation algorithm and as such, three separate investigations were designed.

YouTube

The YouTube experiment aimed to answer two questions: First, whether interaction with extremist content increases the likelihood that users will be shown such content in future, and second, whether the content was ranked higher or lower after interactions. To do this, three identical accounts were created, each following 10 accounts that are identified in the academic literature as being extreme right-wing,[18] and 10 neutral accounts (pertaining to topics such as news, sport or weather). To create a baseline, the accounts did not interact with any content for one week and then were each subjected to a different treatment:

1. The neutral interaction account (NIA) interacted predominantly with neutral content after the first week.[19]
2. The extreme interaction account (EIA) interacted predominately with extreme content after the first week.[20]
3. The baseline account (BA) continued to not interact.

To judge the ranking of content, the authors took the Recommended Videos at the top of the YouTube homepage and ranked content from the top row from left to right (where the top left video is ranked 1, the one directly to its right is ranked 2, and so on). For each session, YouTube offers 18 Recommended Videos, so a video that displayed on the bottom would receive a ranking of 18.[21] Each time an account visited the platform, it watched ten videos in total.

Reddit

The design for the Reddit experiment was similar to that for YouTube. The authors again created three identical accounts that followed 10 identified extreme right-wing sub-reddits and ten neutral sub-reddits (on topics such

18. The authors follow the policy of Vox-Pol and J M Berger by not identifying the names of accounts in this research, both for reasons of potentially increasing exposure and privacy. See J M Berger, *The Alt-Right Twitter Census* (Dublin: Vox-Pol Network of Excellence, 2018).
19. The NIA watched 70% videos from neutral channels and 30% from far-right channels.
20. The EIA watched 70% videos from far-right channels and 30% from neutral channels.
21. These 18 videos represent only a small number of the total recommendations from YouTube, which continue as the user scrolls down the page. However, given that they are in the middle of the homepage, it seems reasonable to assume that they account for a high amount of traffic.

as sports and news). As with the YouTube experiment, a NIA, an EIA and a BA were created. On Reddit, content is ranked from top to bottom, so the top post as ranked as 1, the second as 2, and so on. Reddit displays 25 posts, so the bottom post receives a score of 25.

Gab

Gab is a small, English-language platform that has been identified by scholars as a home to the far-right.[22] The site has a fundamentally different, and more basic, system architecture which calls for a separate approach. Rather than users being recommended posts, there are three different news feeds: Popular (which is the default); Controversial; and Latest. Although Gab does not offer explanations for how these feeds are compiled, the authors consider it most likely that Popular and Controversial are algorithmically driven, at least in part, by non-chronological factors, because these categories require data to reach judgements that are not related to time. They are probably both based on Gab's 'up' and 'down vote' system. Latest is clearly based, either entirely or primarily, on the most recent interactions on the site. This offers a way of comparing the content that is filtered by algorithms against organically posted content. During the project design phase, the authors judged that content does not appear to be reflective of previous interactions or of what other users followed (in other words, it was the same for each user regardless of their interaction history), which would mean the treatment used for YouTube and Reddit would probably return the same results. Instead, data was collected at the same time each day from each of Gab's three news feeds. As with Reddit, content is ranked from top to bottom, so it was judged that Gab ranks in the same way. On Gab, there are four different topics for each news feed: News; Politics; Humour; and Brazil.[23] As none of the researchers possess Portuguese language skills, data was collected from each of the three timelines in the first three topics. Given that no intervention was introduced, this activity is best described as an observation rather than an experiment.

Data collection was undertaken over a two-week period in January/February 2019. For YouTube and Reddit, the authors logged in twice per day,[24] over

22. Lella Nouri and Amy-Louise Watkin, 'Far-Right Hate Group "Britain First" (That Trump Retweeted) Joins Extremist-Friendly Gab', Centre for Analysis of the Radical Right, 3 August 2018, <https://www.radicalrightanalysis.com/2018/08/03/far-right-hate-group-britain-first-that-trump-retweeted-joins-extremist-friendly-gab/>, accessed 10 July 2019.
23. Many supporters of Brazilian President Jair Bolsonaro migrated to Gab after they were banned from Twitter in 2018. See Rodrigo Orihuela, 'Alt-Right Website Gab Attracts Bolsonaro Supporters in Brazil', *Bloomberg*, 4 October 2018.
24. YouTube's Terms of Service mandate that users must not use or launch any automated system that accesses the platform in a manner that sends more request messages to the YouTube servers in a given period of time than a human can reasonably produce. The authors deem accessing the site twice

the 14-day period, creating 28 sessions. This yielded 1,443 videos in total (of which 949 were unique – that is to say, a number of videos appeared in multiple accounts' data collection) on YouTube and 2,100 posts on Reddit (of which 834 were unique). During the data collection period, there were several disruptions to the Gab network, and therefore the authors were only able to log in for five sessions. 1,271 posts were collected, of which 746 were unique, which was a sufficient sample for analysis.

Methodology: Coding[25]

Following collection, the data was coded according to the Extremist Media Index (EMI) developed by Donald Holbrook, which categorises variables as moderate, fringe and extreme.[26] To categorise content as extreme, it must include an incitement to violence or stark dehumanisation. Furthermore, the extreme category is further sub-divided into four levels, based on the specificity of the threat of violence. During the design phase of the research, two members of the team developed and clarified Holbrook's guidance for coding, which can be found in Table 1 in the Appendix.

Results

Having collected and coded the data, the results are presented below. This includes the descriptive data – how much of the content was moderate, fringe or extreme – and the statistical findings as they relate to the research questions.

YouTube

Of the 1,443 videos coded on YouTube, 949 (65.77%) were rated as moderate according to the EMI, while 409 (28.34%) were judged to be fringe and 85 (5.89%) were deemed extremist. Figure 1 shows the distribution of the EMI scores for each session with a rank from one to eighteen, depending on where the video appears on the Recommended Videos section, as well as the percentage distribution of the three categories of content before and after each treatment.[27]

---

per day to be well within what a human can reasonably produce. See YouTube, 'Terms of Service', <https://www.youtube.com/static?gl=GB&template=terms>, accessed 10 July 2019.

25. For more on inter-coder reliability and statistical tests, see the Appendix of this paper.
26. Donald Holbrook, 'Designing and Applying an "Extremist Media Index"', *Perspectives on Terrorism* (Vol. 9, No. 5, 2015), pp. 57–68; Donald Holbrook, 'What Types of Media Do Terrorists Collect?', *International Centre for Counter-Terrorism* (The Hague: The International Centre for Counter-Terrorism, 2017); Donald Holbrook, 'The Terrorism Information Environment: Analysing Terrorists' Selection of Ideological and Facilitative Media', *Terrorism and Political Violence* (14 March 2019), doi: 10.1080/09546553.2019.1583216.
27. In total, 48 videos received a score of zero (mostly because the video had been removed between data collection and coding). These videos are not included in the analysis.

Figure 1: YouTube Data Overview

Does the amount of extreme content increase after applying treatments?

In the EIA, the authors found a statistically significant promotion of both fringe and extreme content: fringe content is 1.37 times more likely,[28] while extreme content in the EIA is 2.00 times more likely than before applying treatment.[29] Furthermore, in the NIA and BA, the amount of extreme content is 2.96[30] and 3.23[31] times less likely, respectively, than the rate of moderate content. That is to say, applying the neutral treatment led to similar results as not applying treatment. These findings suggest that interacting with extreme content on YouTube does increase the prevalence of encountering further extreme content.

Is extreme content better ranked by the algorithm after applying treatments?

In the EIA, before the treatment, there are no observable statistically significant differences in the average rank of content. However, after the treatment, there was a statistically significant difference in rank for extreme and moderate content,[32] although there is no significant difference between fringe and extreme or fringe and moderate. These results suggest that extreme content on YouTube is ranked higher after extreme interaction. There are also no statistically significant differences between EMI scores in the NIA or BA.

Reddit

The data collection from Reddit yielded 2,100 posts, of which 1,654 (78.76%) were moderate, 416 (19.81%) were fringe, while 30 (1.43%) were extreme. Although differences such as platform architecture, different extremist movements,[33] and banning/content removal policy make a direct comparison difficult, the clearest difference between YouTube and Reddit is the difference in the amount of extreme content, with the former having around four times the amount. Figure 2 shows the distribution of the EMI scores for each session with a rank from one to 25, depending on where the content appears on the first page of the timeline as well as the percentage distribution of the three categories of content before and after each treatment.

28. Incident rate ratio (RR) = 1.72, confidence interval (CI) 95% = 1.24–2.37, $p < 0.01$.
29. RR = 2.47, CI 95% = 1.30–4.68, $p < 0.01$.
30. RR = 0.34, CI 95% = 0.12–0.92, $p < 0.05$.
31. RR = 0.31, CI 95% = 0.17–0.58, $p < 0.001$.
32. Extreme Median (Md) = 5, Moderate Md = 10; Wilcoxon Statistic (W) = 1,839, $p < 0.05$.
33. Although there is crossover, the two sites are home to different parts of the extreme right. In the data that was collected for this paper, YouTube videos tended to focus around white supremacy, while Reddit had a higher proportion of men's rights activists.

Figure 2: Reddit Data Overview

Does the amount of extreme content increase after applying treatments?

After estimating the expected frequencies of EMI scores after each treatment, it was found that none of the models for the individual accounts show statistically significant effects. This suggested that interacting with extreme and neutral content on Reddit did not affect the likelihood of a user encountering extreme content.

Is extreme content better ranked by the algorithm after applying treatments?

With regards to the average rank in the EIA, no statistically significant difference in the ranks between any of the variables was found. Similarly, in the NIA, no significant difference could be observed involving extreme content, although interacting with neutral content does appear to decrease the average rank of fringe content,[34] suggesting that there is some filtering in effect on Reddit.

Gab

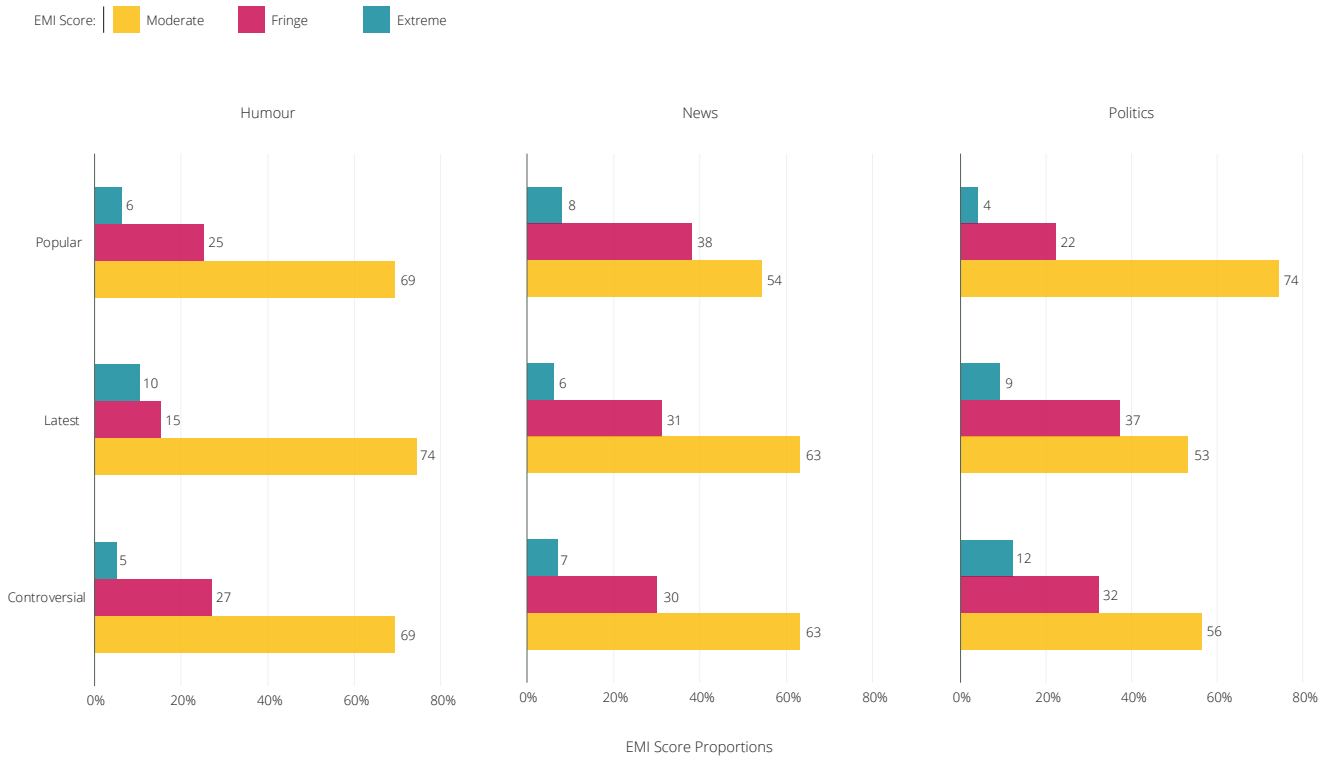
The five visits to Gab returned 1,271 posts, of which 810 (63.73%) were rated as moderate, while 366 (28.8%) were rated as fringe and 95 (7.47%) as extreme. At first glance, these results appear similar to those from YouTube (which yielded almost 6% extreme videos). However, these results must not be misinterpreted given the differences in data collection. For YouTube, accounts were identified that were suspected to be extreme, while on Gab, data was collected on each timeline, regardless of the user.

Is there any difference between the three timelines?

It is worth reiterating that for Gab, the same multiple account treatments were not applied as they were for YouTube and Reddit. Rather, the differences between the three timelines were assessed: Popular (which is the default); Controversial; and Latest within three topics: Humour; News; and Politics. Figure 3 shows the EMI distribution percentages of each of the timelines within each of the topics. The authors observed no statistically significant differences in the EMI scores in the Latest or Controversial timelines in any of the categories. A statistically significant prevalence was found which suggest that fringe content is prioritised above moderate content in the Popular timeline,[35] but not involving extreme content.

34. Moderate Md = 12, Fringe Md = 18, W = 5,576, $p < 0.001$.
35. Moderate Md = 16, Fringe Md = 11, W = 12,494, $p < 0.01$.

Figure 3: Gab Data Overview



Discussion

Evidence was found that only one of the three social media platforms – YouTube – prioritises extreme content following a user's engagement with it. When users do engage with such content on YouTube, they are significantly more likely to be recommended both extreme and fringe content. Similarly, evidence was found that extreme content was pushed up the ranking of recommended videos on the YouTube homepage at the expense of moderate videos. This supports the findings of Derek O'Callaghan and their co-authors, that following the YouTube recommender system while consuming extreme far-right videos may push users into an immersive ideological bubble.[36] It is worth noting that this does not offer an explanation as to whether users' choices play a bigger role in viewers' polarisation than recommender systems, simply that the algorithms can prioritise extreme content.

At the other end of the spectrum, the lack of evidence that Gab's algorithm promotes extreme content is also significant. Gab has been identified as a safe haven for right-wing extremists given its ultra-free speech philosophy,[37] and has been exploited by terrorists such as the Pittsburgh Synagogue attacker Robert Bowers.[38] This suggests that it is the users' choices, rather than a filter bubble effect, that influences this environment. This finding is supported anecdotally, by the project's coders who reported that Gab is by far the most extreme of the platforms they interacted with in the course of conducting this research.

Policy Recommendations

Having established that one of the platforms' recommender systems – YouTube – does promote extremist material after interaction, four policy recommendations are presented. These are not limited to YouTube, but are rather general suggestions for all social media platforms.

36. O'Callaghan et al., 'Down the (White) Rabbit Hole'.
37. Maura Conway, 'Violent Extremism and Terrorism Online in 2018: The Year in Review', Vox-Pol, 2018, <https://www.voxpol.eu/download/vox-pol_publication/Year-in-Review-2018.pdf>, accessed 10 July 2019; Maura Conway with Michael Courtney, 'Violent Extremism and Terrorism Online in 2017: The Year in Review', Vox Pol, 8 December 2017, <https://www.voxpol.eu/download/vox-pol_publication/YiR-2017_Web-Version.pdf>, accessed 10 July 2019; Nouri and Watkin, 'Far-Right Hate Group "Britain First" (That Trump Retweeted) Joins Extremist-Friendly Gab'; Berger, *The Alt-Right Twitter Census*.
38. Alex Hern, 'Gab Forced Offline Following Anti-Semitic Posts by Alleged Pittsburgh Shooter', *The Guardian*, 29 October 2018.

Recommendation 1: Removing problematic content from recommendations

In 2017, Google announced their 'limited features' policy for problematic or controversial videos that do not clearly violate their policies. Such videos will 'appear behind an interstitial warning and they will not be monetised, recommended or eligible for comments or user endorsements'.[39] Despite this policy, the authors found that 6% of recommended videos were extreme, and much of the fringe content came from accounts that were also producing extreme content. More broadly, this approach is similar to Reddit's 'quarantine' system, where users must opt-in to view potentially offensive or upsetting content. In this system, no advertising revenue can be gained and content does not appear in non-subscription-based feeds.[40] This type of system may be an effective model for other social media platforms to find a constructive balance between freedom of speech and harmful content.

Recommendation 2: Ensuring recommendations are from quality sources, and providing users with more context and alternative perspectives

In their steps to counter disinformation online, Google has introduced changes to deliver information from quality, trustworthy sources and to give users more context in results from searches and Google News[41]. The authors suggest a similar approach to the selection of recommended videos. For example, for users viewing extreme far-right content, recommended videos should come from high-quality and highly credible sources (such as trusted news outlets with rigorous fact-checking). Similarly, efforts to include voices that challenge extremism from within a potential filter bubble, such as Jigsaw's redirect method, should be applied to all platforms.[42] These alternative voices should be of the highest factual quality.

39. Kent Walker, 'Four Steps We're Taking Today to Fight Terrorism Online', Google, 18 June 2017, <https://www.blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/>, accessed 10 July 2019.
40. Reddit, 'Content Policy Update', 2015, <https://www.reddit.com/r/announcements/comments/3fx2au/content_policy_update/>, accessed 10 July 2019.
41. Google, 'How Google Fights Disinformation', February 2019, <https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Disinformation.pdf>, accessed 10 July 2019.
42. The redirect method creates banner advertisements when users type potentially problematic language into Google, which gives the user an opportunity to click and be redirected to a YouTube playlist of curated strategic communications to dissuade them from the ideology. See, The Redirect Method, 'About the Method', <https://redirectmethod.org/>, accessed 10 July 2019 .

Recommendation 3: Greater transparency

Social media platforms provide users with a clear option to request why content has been recommended to them. Facebook operates a policy for advertisements which offers users an opportunity to click on an options menu to find out 'why am I seeing this ad?'[43] which gives users a short explanation of why it has been recommended. In April 2019, Facebook announced that it will launch a new feature that explains why its algorithm has displayed this content.[44] Other social media platforms should consider this approach for content recommendations. This option should be as clearly marked and as easily discoverable as possible. This is in keeping with the growing academic movement towards 'explainable artificial intelligence', which offer descriptions of how and why automation affects user experience.[45]

Recommendation 4: Further research

This research project was constrained terms of the number and type of social media platforms that could be researched, either by the closed nature of platforms (such as Facebook) or by the restrictions within the terms of service (such as Twitter). Maura Conway observes the problem of studying violent extremism is that certain sites are easier for researchers to access, creating a body of knowledge for those sites while leaving large gaps for others.[46] Many of the questions identified in previous research, such as how extremists' behaviour is affected by personalisation algorithms (rather than the content they see),[47] or the effects of non-optional algorithms, such as news feeds or timelines,[48] are left unanswered. To be able to address these questions, the authors recommend greater collaboration between social media platforms and researchers.

43. Facebook Help Centre, 'Why am I Seeing Ads From a Specific Business or Advertiser?', <https://www.facebook.com/help/1674984446161704>, accessed 10 July 2019.
44. *BBC*, 'Facebook to Reveal News Feed Algorithm Secrets', 1 April 2019.
45. Derek Doran, Sarah Schulz and Tarek R Besold, 'What Does Explainable AI Really Mean? A New Conceptualization of Perspectives', <https://arxiv.org/pdf/1710.00794.pdf>, accessed 10 July 2019 .
46. Maura Conway, 'Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research', *Studies in Conflict & Terrorism* (Vol. 40, No. 1, 2017), pp. 77–98.
47. O'Callaghan et al., 'Down the (White) Rabbit Hole'.
48. Bakshy, Messing and Adamic, 'Exposure to Ideologically Diverse News and Opinion on Facebook'.

Alastair Reed is an Associate Professor at Swansea University; he is also an Associate Professor at TU Delft in the Netherlands.

Joe Whittaker is a doctoral student at Swansea University and Leiden University. He is also a research fellow at the International Centre for Counter-Terrorism.

Fabio Votta is a graduate student in Empirical Political and Social Research at the University of Stuttgart.

Seán Looney is a doctoral student at Swansea University and Université Grenoble Alpes.

Methodological Appendix

Table 1: Holbrook's Extremist Media Index,[49] with definitional guidance. Holbrook's original remarks are reproduced here, with the authors' clarifications in italics.

Primary Grading Category	Definition
1 – Moderate	General religious, political, philosophical or historical material and news commentary containing no endorsement of violence or hatred towards identified communities with generally moderate content along the lines found in mainstream religious/political texts and news media output. *This can include reference to an out-group if it is something that is regarded as part of acceptable political discourse, such as criticising a political party for causing problems.*
2 – Fringe	Content is religiously or ideologically conservative and isolationist, politically radical and confrontational, but without any justifications conveyed for violence in present-day scenarios. Anger and hostility might be expressed towards a given group of people, such as the 'kuffar' (unbelievers) or immigrants, without the added assumption that these people are somehow 'subhuman' and legitimate targets of violence. *This can include:* • *Profanity laden nicknames for the out-group that go beyond political discourse (such as 'libtards' or 'feminazis').* • *Political speech that goes beyond political norms (such as 'Democrats are traitors').* • *Historical revisionism of a settled or complicated issue to blame an out-group (but without justification of violence or stark dehumanisation).* • *Explicit support for other fringe/extremist movements (again with no reference to violence or dehumanisation).* • *Prescription for ordering society based on historical structures (for example, 'men have always protected women, therefore they still should').*
3 – Extreme	Material that legitimises and/or glorifies the use of violence, especially serious and potentially fatal violence, to achieve particular goals, as well as the fighters and martyrs who die for the cause, with some allusion to the view that such prescriptions continue to be relevant for contemporary activists. Also included within this category is material that focuses on dehumanising particular communities, citing issues of race, sexuality, origin or other aspects that render such people 'sub-human,' thus undermining their right to life. This category thus captures both publications advocating 'jihadi' violence against combatants or civilians, as well some works of the extreme right-wing, for instance, that can be more opaque in terms of references to violence but with a focus on presenting people such as Jews and non-whites as sub-human in the context of an imagined or envisaged confrontation with these groups of people.
Secondary Grading Category	**Definition**
Extreme Level 1	Serious violence (i.e. potentially fatal) is only justified/promoted/welcomed with reference to combatants or is vague, without any detail, e.g. talk about the virtues of collective violence, glorification of insurgency warfare.
Extreme Level 2	Serious violence (i.e. potentially fatal) clearly justified/ promoted/welcomed against non-combatants, but without any detail, e.g. "murder Muslims", "kill the kuffar".
Extreme Level 3	Serious violence (i.e. potentially fatal) justified/promoted/ welcomed against non-combatants and with some detail regarding facilitation, scope or direction: i.e. "do suicide attacks" (against non-combatants), "target the economy".
Extreme Level 3b	Same as '3' but specific and directly applicable details offered, e.g. bomb-making recipes.

49. Holbrook, 'Designing and Applying an "Extremist Media Index"'.

Coding

To ensure inter-coder reliability, two coders worked on a random sample of 35 pieces of the same content from each site (for a total of 105). The overall agreement was 80.76% (YouTube = 74.3%, Reddit = 85.7%, Gab = 81.8%), yielding a Krippendoff's alpha value of 0.74 (YouTube = 0.77, Reddit = 0.72, Gab = 0.73). Although this is below the ideal value of $\alpha \geq 0.800$, the authors accept tentative conclusions as the reliability is well above the lowest conceivable limit of 0.667. This is also higher than in Holbrook's research.[50]

The same two coders then coded the remaining content on each platform. This included only the original content, namely the YouTube video, the original Reddit post, or the original Gab post. Comments underneath were not taken into account. For YouTube videos, only the first five minutes were viewed, as many of the videos were several hours long. While imposing a time limit is arbitrary and undesirable, it is the only way that every video could be coded, without having to exclude some for being too long, which is more undesirable as the sample would not be representative. Finally, any links to other sites were not included in the coding.

Tests

Quasi-Poisson models were used to estimate rate ratios and expected frequency counts to test whether extremist or fringe content was more or less prevalent after treatments were applied.[51] To test for rank differences in content, Wilcoxon rank sum tests were chosen, a non-parametric alternative to the unpaired two-samples t-test. This method was chosen because the data is found to violate the assumption of having a normal distribution and the Wilcoxon rank sum tests account for that by comparing median values instead of arithmetic means.

50. *Ibid.*
51. Alan Agresti, *Categorical Data Analysis*, 3rd Edition (Hoboken, NJ: John Wiley & Sons, 2013).