Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence

1 author:
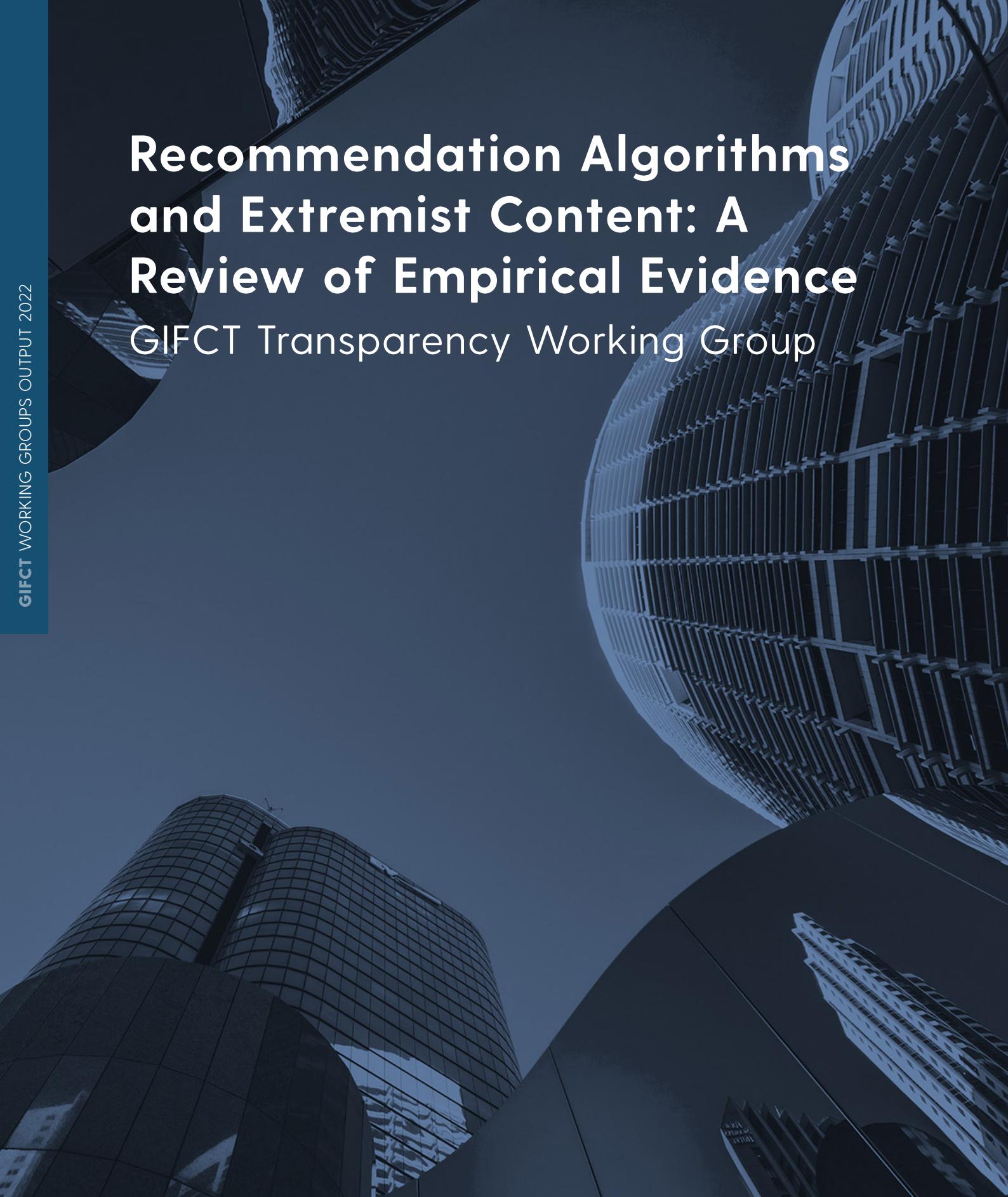
Joe Whittaker
Swansea University
26 PUBLICATIONS **299** CITATIONS

Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence

GIFCT Transparency Working Group

GIFCT
Global Internet Forum
to Counter Terrorism

Dr. Joe Whittaker
Swansea University

Introduction

This paper reviews the existing empirical studies on the role of social media recommendation algorithms and potential links to extremist content. It seeks to provide transparency for future researchers by taking stock of the present empirical knowledge base, noting the types of data and methods that are utilized and charting gaps in the research. In doing so, the paper sheds light on definitional issues, replicability, agency, causality, and the limitations of present approaches. It also offers a window into how social media companies have adapted their practices over the past decade by surveying public statements about their policies. In total, 15 studies were identified for review. The appendix contains a table outlining each study's methods and findings.

Out of the review, a nuanced picture emerges of research into the role of recommendation systems and extremist content. There is a heavy emphasis on studying YouTube, a focus on the far-right, as well as English-language content. Moreover, many of the studies collect and analyze data in the mid-2010s, which importantly was before many platforms began to downrank or remove borderline content from recommendations. Although there is a wide array of methods utilized to investigate this topic, all but two papers rely on an external "black box" methodology in which researchers cannot manipulate platforms' recommendation systems. Similarly, there are only three studies in total which utilize a control group. More often, researchers access a platform's Application Programming Interface (API) to establish content that could potentially be recommended, which cannot account for personalization. There are also substantial differences in how "extremism" or related concepts are deployed in coding systems, which is a challenge for meta-reviews such as this. Most of the studies show that platforms can recommend extremist content, although there are several important caveats, such as the importance of pre-existing user beliefs and related variables. This review concludes by contextualizing the findings in the wider academic and policy debates, as well as offering a set of recommendations moving forward.

Before beginning, it is worthwhile to be clear about the topic under investigation. The Global Internet Forum to Counter Terrorism's (GIFCT) Content-Sharing Algorithms, Processes, and Positive Interventions Working Group highlights three categories of social media algorithms that could be exploited by violent extremists:

  i. **Search algorithms** – such as autocompleting a keyword;
  ii. **Recommendation algorithms** – which curate content that a user may be interested in; and
  iii. **Ad-tech algorithms** – that target users based on demographics and behavior to optimize advertising.[1]

This review is focused specifically on recommendation algorithms, which Ricci, Rokach, and Shapira define as software tools and techniques which provide users with suggestions for items a user may

1 "Content-Sharing Algorithms, Processes, and Positive Interventions Working Group," Global Internet Forum to Counter Terrorism, 2021, https://gifct.org/wp-content/uploads/2021/07/GIFCT-CAPI1-2021.pdf.

wish to utilize.[2] An important distinction is whether the content is pre-selected or user-selected.[3] For example, YouTube's "Recommended for you" algorithm offers algorithmically driven suggestions that the user has the option to select or ignore; on the other hand, platforms with news feeds or timelines do not necessarily offer this choice, with content appearing before a user has selected it.[4] This review follows the lead of the above-mentioned GIFCT report in including both pre-selected and user-selected recommendation systems – i.e. inclusive of timelines, news feeds and recommended videos, etc.[5]

The scope of this review spans studies that analyze extremist content. However, "extremist" is an essentially contested concept,[6] as are related words such as "terrorist" and "radical," which are often used interchangeably, and which have led to considerable conceptual ambiguity.[7] Depending on one's conceptualization, extremist content on social media may include terrorist propaganda, materials that directly advocate violence, incitement of hatred, and/or non-illegal yet potentially harmful content that may "other" certain out-groups. To be included in this review, the authors of a study must identify the content as extremist, radical, or terrorist, or refer to specific ideologies that are widely considered to be extreme (such as the far-right or jihadism). While this is an imperfect criterion, it is worthwhile to be inclusive, and then utilize this review to analyze the decision-making of the researchers and coding systems in the corpus of literature to determine what kind of content is deemed extreme.

Given this scope, this review does not include research into the wider field of political discourse and polarization. Therefore, studies such as Bakshy, Messing, and Adamic's research into Facebook's News Feed or Cho and colleagues' laboratory experiment of YouTube's recommendation system are omitted.[8] Although there is a link between polarization and extremism,[9] as well as concerns over the normalization of far-right narratives into mainstream politics,[10] this review is focused specifically on content that has been explicitly identified as extreme. Similarly, there is a growing empirical literature

• •

2 Francesco Ricci, Lior Rokach, and Bracha Shapira, Recommender Systems Handbook, (New York: Springer, 2011).

3 Frederik J. Zuiderveen Borgesius et al., "Should we worry about filter bubbles?" Internet Policy Review 5, no. 1 (March 31, 2016). https://doi.org/10.14763/201 6.1.401.

4 Joe Whittaker et al., "Recommender Systems and the Amplification of Extremist Content," Internet Policy Review, 10, no. 2 (June 30, 2021), https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content.

5 "Content-Sharing Algorithms," GIFCT.

6 Walter Bryce Gallie, "Essentially Contested Concepts," Proceedings of the Aristotelian Society, 56, (1955): 167–198.

7 For example, see Randy Borum, "The Etiology of Radicalization," In The Handbook of the Criminology of Terrorism, eds. G. LaFree and J. D. Freilich (Chichester: John Wiley and Sons, 2017): 17–32; Bart Schuurman and Max Taylor, "Reconsidering Radicalization: Fanaticism and the Link Between Ideas and Violence," Perspectives on Terrorism, 12, no. 1 (2018): 3–22.

8 Eytan Bakshy, Solomon Messing and Lada A. Adamic, "Exposure to Ideologically Diverse News and Opinion on Facebook," Science Express, (May, 2015), 1–5, doi: 10.1111/j.1460-2466.2008.00410.x.; Jaeho Cho et al., "Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization," Journal of Broadcasting and Electronic Media 64, no. 2 (2020): 150–172.

9 Cass R. Sunstein, "The law of group polarization," The Journal of Political Philosophy, 10, no. 2, (2002): 175–195; Bertjan Doosje et al., "Terrorism, Radicalization and De-radicalization," Current Opinion in Psychology 11 (2016): 79–84.

10 Aurelien Mondon and Aaron Winter, "Articulations of Islamophobia: From the extreme to the mainstream?," Ethnic and Racial Studies, 40, no. 13 (2017): 2151–2179.

on recommendation systems and disinformation or misinformation.[11] While there is an overlap between disinformation and extremism,[12] only studies that are explicitly related to extremism have been included. One study was excluded despite making specific reference to "extreme content" in the title as the study did not actually analyze extremism but instead focused on "contextually inappropriate" recommendations such as satire and sexually suggestive material being found by children.[13]

Data

YouTube is the most popular platform for empirical research within the existing literature; nine of the studies analyzed whether its recommendation system promotes extreme content. This is followed by Twitter, whose recommendations or timeline are analyzed in three studies. Two studies explore Reddit, either looking at the up and downvoting system or its "Best" timeline, with one study each looking at Facebook and Gab. Finally, one study is interview-based and does not explicitly focus on one platform, although the findings mention YouTube's recommendation system. YouTube's dominance in the literature is highlighted by Whittaker and colleagues, who note that it has a researcher-friendly API compared to other social media platforms; they suggest that the field may be driven by research convenience rather than necessarily following the trail of extreme content.[14]

The studies are weighed towards researching far-right content. This is counter to the general pattern in the wider field of terrorism and extremism studies, which has often been noted as being primarily focused on jihadism.[15] Six of the studies examine far-right content exclusively, while three others study the far-right and another ideology (such as jihadism, male supremacism, or the far-left). Five focus exclusively on jihadism or Islamism, while one analyses the incel community. The emphasis on the far-right may be related to the relative success with which social media platforms have been able to identify and remove jihadist content and the difficulties in applying the same lessons to far-right content.[16] In other words, researchers may be choosing to use far-right data because they have not yet been removed.

Perhaps owing in part to the focus on the far-right, there is also an English-speaking and Western

11 For example, see Eslam Hussein, Prerna Juneja, and Tanushree Mitra, "Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube," Proceedings of the ACM on Human-Computer Interaction, 4 (2020), doi: 10.1145/3392854; Konstantinos Papadamou et al., "'It is just a flu': Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations," (2020), http://arxiv.org/abs/2010.11638; Jonas Kaiser, Adrian Rauchfleisch, and Yasodara Córdova, "Fighting Zika With Honey: An Analysis of YouTube's Video Recommendations on Brazilian YouTube," International Journal of Communication 15, no. 108 (2021): 1244–1262.

12 Jamie Bartlett and Carl Miller, "The Edge of Violence: Towards Telling the Difference Between Violent and Non-Violent Radicalization," Terrorism and Political Violence, 24, no. 1 (2012), 1–21; Jacob Davey and Julia Ebner, "The Great Replacement: The Violent Consequences of Mainstreamed Extremism," Institute for Strategic Dialogue (2019), https://www.isdglobal.org/wp-content/uploads/2019/07/The-Great-Replacement-The-Violent-Consequences-of-Mainstreamed-Extremism-by-ISD.pdf.

13 Christian Stöcker and Mike Preuss, "Riding the Wave of Misclassification: How we end up with extreme YouTube content," Lecture Notes in Computer Science (July 10, 2020): 359–375.

14 Whittaker et al., "Recommender Systems."

15 Maura Conway, "Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research," Studies in Conflict and Terrorism, 40, no. 1 (2017): 77–98; Bart Schuurman, "Topics in Terrorism Research: Reviewing trends and gaps, 2007–2016," Critical Studies on Terrorism, 12, no. 3 (2019): 463–480.

16 Maura Conway, "Routing the Extreme Right: Challenges for Social Media Platforms," RUSI Journal 165, no. 1 (2020): 108–113.

focus to the data in existing studies. In seven cases, the "seed" accounts that are utilized to begin data collection identify English-speaking accounts either exclusively or predominantly. German content is also utilized; twice in studies with one English and one German-speaking dataset, and once exclusively in German. Two studies utilize Arabic-language content, and three studies had a mixed set of languages which include English, Japanese, French, Spanish, German, Turkish, Arabic, and Mandarin.

Research Within a Changing Social Media Landscape

A relevant factor within the corpus of academic literature is when the data for each individual study are collected. Two studies collected data in 2013,[17] Murthy's research uses a dataset from 2016,[18] with two more in 2017[19] and one study spanning October 2017 to March 2018.[20] Ledwich and Zaitsev's study does not explicitly state the date of data collection, but it can be inferred that it likely took place in 2019.[21] Four further studies use data from that year,[22] with one study doing so in 2020.[23] Two take longitudinal approaches that span multiple years: Hosseinmardi et al. collect viewing behaviors from January 2016 to December 2019 and Huszár et al. access timeline data for users from June 2016 to June 2020.[24] The article by Baugut and Neumann does not specify dates for reasons of anonymity, but the interviews were conducted a relatively short time before publication.[25] As this study involved participants reflecting on their propaganda use in the past, it is difficult to assign a general timeframe.

Understanding when these studies collected data is important because the landscape of social media regulation has changed dramatically in the past decade. Academics point to 2016 as a turning point in which platforms began to take a more proactive approach toward removing content and

17 Derek O'Callaghan et al., "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems," Social Science Computer Review 33, no. 4 (2015), 459–478; J. M. Berger, "Zero Degrees of al Qaeda," Foreign Policy, August 14, 2013, http://foreignpolicy.com/2013/08/14/zero-degrees-of-al-qaeda/.

18 Dhiraj Murthy, "Evaluating Platform Accountability: Terrorist Content on YouTube," American Behavioral Scientist, 65, no. 6 (2021): 800–824.

19 Josephine B. Schmitt et al., "Counter-messages as prevention or promotion of extremism?! The potential role of YouTube," Journal of Communication 68, no. 4 (2018): 758–779; Tiana Gaudette, Ryan Scrivens, and Garth Davies, "Upvoting Extremism: Collective identity formation and the extreme right on Reddit," New Media and Society (September, 2020), doi: 10.1177/1461444820958123.

20 Gregory Waters and Robert Postings, "Spiders of the Caliphate: Mapping the Islamic State's Global Support Network on Facebook," Counter-Extremism Project (May, 2018).

21 Mark Ledwich and Anna Zaitsev, "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization," Eprint arXiv:1912.11211, (2019).

22 Manoel H. Ribeiro et al., "Auditing Radicalization Pathways on YouTube," Woodstock '18: ACM Symposium on Neural Gaze Detection (2019) http://arxiv.org/abs/1908.08313; Michael Wolfowicz, David Weisburd and Badi Hasisi, "Examining the interactive effects of the filter bubble and the echo chamber on radicalization," Journal of Experimental Criminology (2021), doi:10.1007/s11292-021-09471-0.1; Whittaker et al., "Recommender Systems,"; Kostantinos Papadamou et al., "How over is it?" Understanding the Incel Community on YouTube," Proceedings of the ACM on Human-Computer Interaction 5 (2021), https://arxiv.org/abs/2001.08293.

23 Annie Y. Chen et al., "Exposure to Alternative and Extremist Content on YouTube," Anti-Defamation League, https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube.

24 Homa Hosseinmardi et al., "Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube" (2020), arXiv; Ferenc Huszár et al., "Algorithmic Amplification of Politics on Twitter," Proceedings for the National Academy of Sciences of the United States of America, 119, no. 1 (2022).

25 Information provided by Dr. Philip Baugut by email; see Philip Baugut and Katharina Neumann, "Online propaganda use during Islamist radicalization," Information Communication and Society, 23, no. 11 (2020): 1570–1592.

disrupting accounts sympathetic to terrorism.[26] Prior to this, sites often framed themselves as allowing free speech to run its course on their sites. In 2012, Twitter's then-General Manager Tony Wang noted that: "We remain neutral as to the content because our general counsel and CEO like to say that we are the free speech wing of the free speech party."[27] Similarly, YouTube's decision to remove non-violent videos of Anwar al-Awlaki inciting violence in 2017 represented a key policy change, which saw a more proactive approach towards terrorist content.[28] This does not mean that terrorist content was not removed prior to 2016, but that platforms have become more proactive and as a result more sophisticated at detecting and removing content. In short, early studies in the cohort may draw from data that would not be recommended today. Murthy makes this point explicitly, noting that his dataset (derived from 2016) pre-dates the formation of GIFCT, which signaled increased efforts to prevent terrorists and violent extremists from exploiting digital platforms.[29]

While content removal is an important factor – extremist material cannot be recommended if it is not available on the platform – given the scope of this review, it is worthwhile to consider how platforms have changed their approach to recommendation systems when faced with potentially problematic content that does not clearly violate some aspects of platforms' rules or terms of service. Below, an overview of policy for each of the five different platforms under study (YouTube, Facebook, Twitter, Reddit, and Gab) within the corpus of literature is presented to ascertain whether (and if so how) they changed their recommendations in the face of extreme content.

The first platform to publicly announce an alteration to recommendations was Reddit, which introduced a policy of "quarantining" subreddits in 2015. When the platform applies these measures, the subreddit is only viewable to those who explicitly opt-in.[30] This approach is taken to "prevent its content from being accidentally viewed by those who do not knowingly wish to do so, or viewed without appropriate context."[31] Quarantined subreddits do not appear in non-subscription-based feeds (such as Reddit's "Popular" feed) and are not included in search or recommendations. This is relevant for both studies on Reddit in this corpus: Gaudette and colleagues collect data from r/The_Donald in 2017, which was subsequently quarantined in 2019 before eventually being banned.[32] On the other hand, Whittaker et al.'s study found that Reddit's "Best" timeline did not recommend extreme content, which could be a result of having removed problematic content due to quarantining.[33]

YouTube takes a four-pronged approach to content moderation: Removing problematic and violative

26 J. M. Berger and Heather Perez, "The Islamic State's diminishing returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters," George Washington University: Program on Extremism, (February 2016); Maura Conway, "Violent Extremism and Terrorism Online in 2016: The Year in Review," Vox Pol (2016).

27 Josh Halliday, "Twitter's Tony Wang: 'We are the free speech wing of the free speech party,'" The Guardian March 22, 2012, https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

28 Scott Shane, "In 'Watershed Moment,' YouTube Blocks Extremist Cleric's Message," New York Times, November 27, 2017, https://www.nytimes.com/2017/11/12/us/politics/youtube-terrorism-anwar-al-awlaki.html.

29 Murthy, "Evaluating Platform Accountability."

30 Reddit, Content Policy Update, (2015), https://www.reddit.com/r/announcements/comments/3fx2au/content_policy_update/.

31 Reddit, Quarantined Subreddits, (2021), https://www.reddithelp.com/hc/en-us/articles/360043069012.

32 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

33 Whittaker et al., "Recommender Systems."

content from the platform; Raising up authoritative voices; Rewarding trusted creators; and Reducing the recommendations of borderline content.[34] The tactic of Reducing was first articulated in 2017 as a counter-terrorism policy, noting that the platform would take a "tougher stance on videos that do not clearly violate policies but may be inflammatory or supremacist" by removing content from being recommended.[35] According to YouTube, this step reduced views of such videos by an average of 80%.[36] This tactic was expanded to misinformation and conspiracy theories in early 2019 (YouTube 2019c),[37] which saw a drop of 70% in views of this content.[38] In mid-2019, the platform announced that they were expanding this policy by including authoritative voices into potential recommendations when an individual is watching borderline content.[39] YouTube also reward their trusted creators financially through their monetization program.[40] Although YouTube's policy has been updated several times, several studies collected data before the initial changes in 2017 (for example, the studies by O'Callaghan et al., Schmitt et al., and Murthy) while others did so after the first policy but before more recent updates, such as the public articulation of the 4Rs framework, including the research by Ledwich & Zaitsev, Ribeiro et al., and Whittaker et al., with the study by Chen and colleagues taking place after all of the updates outlined above. The data collected in the longitudinal studies (i.e. Hosseinmardi et al. and Papadamou et al.) reflects several changes in recommendation policy.

Facebook takes an approach similar to YouTube's, demarcating their moderation policy into three prongs: Removing violative content, Reducing misleading content via ranking and Informing users with additional context. This policy was outlined in 2018, mostly framed around sensationalist material and clickbait, noting that problematic content that does not violate policies can still be harmful to users, and when identified was downranked in the platform's News Feed.[41] However, Facebook's policy has since been updated to explicitly include content that may incite hatred, particularly in countries at risk of conflict.[42] Facebook also operates a Dangerous Individuals and Organizations policy, which seeks to restrict recommending movements that may be tied to violence but do not meet the criteria to be banned. The pages of groups designated as such by this policy are not eligible to be recommended

---

34 "The Four Rs of Responsibility, Part 1: Removing harmful content," YouTube (September 3, 2019) https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/.

35 Kent Walker, "Four steps we're taking today to fight terrorism online," Google, June 18, 2017, https://blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/.

36 "Our Ongoing Work to Tackle Hate," YouTube, June 5, 2019, https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/.

37 "Continuing Our Work to Improve Recommendations on YouTube," YouTube, January 25, 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve/.

38 "Managing Harmful Conspiracy Theories on YouTube," YouTube, October 15, 2020, https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube/.

39 "Our Ongoing Work."

40 Neal Mohan, "Responsibility is good for business and for the economy," YouTube (Blog), August 23, 2021, https://blog.youtube/inside-youtube/responsibility-good-business-and-creator-economy/

41 Tessa Lyons, "The Three-Part Recipe for Cleaning up Your News Feed," Facebook Newsroom, May 22, 2018, https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/; Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement," Facebook, November 15, 2018, https://www.facebook.com/notes/751449002072082/.

42 Samidh Chakrabarti and Rosa Birch, "Understanding Social Media and Conflict," Facebook Newsroom, June 20, 2019, https://about.fb.com/news/2019/06/social-media-and-conflict/.

and are downranked in the News Feed, as well as not suggested in the search function.[43] Only one study in the corpus studies Facebook and its findings relate to friend recommendations, so it is unclear whether its data is implicated by any of these policies.[44]

Twitter updated its Hateful Conduct Policy in 2019 to include several different enforcement options for dealing with hate speech.[45] This included existing consequences such as account suspension and the removal of tweets, but also the ability to downrank tweets within replies, making tweets ineligible for amplification in "Top Search" and/or on timelines for users that do not follow the author and excluding tweets and accounts in email or in-product recommendations.[46] Each of the three studies in this corpus began data collection before this policy change, although one study is longitudinal and the new policy was implemented during its time frame.[47]

Finally, Gab does not appear to have a policy that removes content from their recommendations (such as their "Popular" timeline). Gab claims that they use the First Amendment as their guiding principle and "make the best effort to ensure that all content moderation decisions and enforcement... does not punish users for exercising their... right to speak freely".[48] They state that they reserve the right to remove content or ban accounts that they feel violate the First Amendment's protection, but do not mention any kind of algorithmic downranking.

Methods

Research Objective

The role of recommendation systems is not the primary research objective or dependent variable in every study in the corpus. It is the main focus in seven,[49] for three it was just one of several research questions,[50] while for two, it is tested alongside another variable, such as proxies for an "echo chamber" effect.[51] The research of Huszár and colleagues is primarily concerned with recommendation algorithms, but the proliferation of far-right and far-left accounts is only one of several research questions.[52] Some studies had other goals, which led to findings that offer a perspective on algorithms. Waters and Postings conduct an analysis of online supporters and had incidental findings that relate to Facebook's

43 "An Update to How We Address Movements and Organizations Tied to Violence," Facebook Newsroom, August 19, 2020, https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

44 Waters and Postings, "Spiders of the Caliphate."

45 Twitter, "Updating our Rules Against Hateful Conduct," July 9, 2019, https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.

46 Twitter, "Hateful Conduct Policy," https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

47 Huszár et al., "Algorithmic Amplification."

48 Gab, Website Terms of Service, April 10, 2020, https://gab.com/about/tos.

49 Berger, "Zero Degrees"; O'Callaghan et al., "Down the (White) Rabbit Hole"; Schmitt et al., "Counter-messages as prevention"; Ledwich and Zaitsev, "Algorithmic Extremism"; Gaudette, Scrivens, and Davies, "Upvoting Extremism"; Whittaker et al., "Recommender Systems"; Murthy, "Evaluating Platform Accountability."

50 Ribeiro et al., "Auditing Radicalization Pathways"; Papadamou et al., "How over is it?"; Chen et al., "Exposure to Alternative & Extremist Content."

51 Hosseinmardi et al., "Evaluating"; Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

52 Huszár et al., "Algorithmic Amplification."

recommendation system.[53] Similarly, Baugut and Neumann seek to understand the media diet of radical Islamists, who in turn self-report the importance of social media algorithms.[54]

Internal Versus External Access

A key distinction is how researchers accessed the respective datasets. In their report on responsible AI for social media, Knott and colleagues dichotomize between "Internal" access – i.e. studies in which social media companies work with researchers and provide private privileged data access – and "External" access – which uses publicly available datasets.[55] In this review, fourteen of the studies accessed external data, while only one relied on internal access: Huszár et al.'s study involved internal Twitter data and a research team comprising platform staff and academics.[56] The vast majority of studies in this corpus rely on "black box" testing in which researchers input data and receive outputs without having an understanding of how the underlying algorithms makes decisions.[57] Knott et al. argue that this discrepancy is a key limitation of the academic body of knowledge; external studies do not test causal hypotheses about the effects of recommender systems. Some of these types of studies, they argue, can manipulate users (either real or automated), but none of them can manipulate the platforms' recommendation system to observe its effect on users.[58] This has important ramifications: it is difficult to achieve algorithmic transparency given the "black box" nature of platforms' recommendation systems, which are closely guarded trade secrets as exposing their inner workings may lead to a competitive disadvantage[59] as well as opening them to gaming from bad actors.[60]

Experimental

Three studies seek to create an experimental condition that test a treatment against a control group. Wolfowicz et al. conduct a study of 96 young males in East Jerusalem that, prior to the study, did not use Twitter.[61] They test whether there was an interactive relationship between filter bubbles, echo chambers, and the justification of violence. The treatment group suppressed algorithms by signing up to a new Twitter account with a new email and rejected all the platforms' automated recommendations. The control group used existing emails and accepted the recommendations. They tested for an echo chamber effect using a range of social network variables. As well as using data from Twitter, they also asked the individuals survey questions, such as whether they felt that suicide bombings were ever justified. This can be considered an example of an externally accessed study

53 Waters and Postings, "Spiders of the Caliphate."

54 Baugut and Neumann, "Online propaganda use."

55 Alistair Knott et al., "Responsible AI for Social Media," The Global Partnership on Artificial Intelligence, (2021), https://gpai.ai/projects/responsible-ai/social-media-governance/responsible-ai-for-social-media-governance.pdf.

56 Huszár et al., "Algorithmic Amplification."

57 "Content-Sharing Algorithms," GIFCT.

58 Alistair Knott et al., "Responsible AI for Social Media."

59 Michael A. DeVito, "From Editors to Algorithms," Digital Journalism, 5, no. 6 (2017): 753–773.

60 Nicholas Diakopoulos, "Algorithmic Accountability: Journalistic investigation of computational power structures," Digital Journalism, 3, no. 3 (2015): 398–415.

61 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

that can manipulate the recommendation algorithm to test the potentially harmful effects on users, although it still does not have access to the inner workings of the algorithm.

Whittaker et al. create experimental conditions on YouTube and Reddit by creating three accounts and following the same set of channels or subreddits (10 far-right; 10 neutral).[62] The accounts were then left dormant for a week so the recommendations could be collected twice per day without any interaction to create a baseline. Then, each of the accounts was subjected to one treatment: one interacted primarily with far-right channels or content, one interacted primarily with neutral content and one continued to do nothing. Each of these treatments also lasted for one week. This offered a basis of comparison both against the baseline (i.e. what has changed from the beginning?) and the control groups (i.e. how has interaction with far-right content changed compared to other alternatives?).

Finally, Huszár et al. create a randomized natural experiment on Twitter.[63] When Twitter introduced a machine-learning personalized timeline in 2016, it excluded 1% of its users, who instead see content in chronological order. This latter group acted as a control, allowing the researchers to compare whether certain types of politicians (including far-right and far-left) or media source had greater algorithmic amplification (i.e. if their tweets reached a higher proportion of personalized timelines than chronological ones).

Tracking User Behaviors

Two studies track how users act online. Attempting to understand echo chambers and recommendations on YouTube, Hosseinmardi et al. use longitudinal data from a US nationally-representative sample of over 300,000 users from Nielsen's desktop web panel from 2016 until the end of 2019.[64] Their research objective is to investigate whether YouTube's recommendation system systematically directs users to far-right content. To do this, they examined how much of overall consumption was made up of news-related content; whether far-right channels had increased over the research period; the pathways toward far-right channels (i.e. from recommendations or not); and whether longer session times led to more extreme content.

Chen and colleagues use a similar approach in coordination with a nationally-representative survey.[65] Firstly, participants conducted the Cooperative Congressional Election Survey, which asked a range of questions about politics, society, and values. Then, a sample of the participants was asked to install an extension that tracks their activity on YouTube (including which recommendations were shown and engaged with). The combination of survey and tracking data allowed the researchers to explore whether individuals with existing beliefs – such as racial resentment – were more likely to be shown extreme content than those who do not hold such beliefs.

62 Whittaker et al., "Recommender Systems."

63 Huszár et al., "Algorithmic Amplification."

64 Hosseinmardi et al., "Evaluating."

65 Chen et al., "Exposure to Alternative & Extremist Content."

API-Mining

The most common approach employed by the studies under review is accessing a platform's Application Programming Interface (API) to view content that could be shown as part of recommendations. O'Callaghan et al. utilize a set of seed channels on YouTube to acquire metadata for 1000 videos, which were randomly sampled for up to 50 videos per seed.[66] They then acquired metadata for the top ten Related Videos (i.e. videos that would be recommended). The metadata for the channels of these videos were then retrieved in the same way – metadata for 1000 videos and then up to 50 were randomly sampled. These data were then analyzed to establish the extent to which the related channels for a far-right seed feature extremist content.

Ribeiro et al. also use the YouTube API to conduct a large-scale audit of what they call "user radicalization" by identifying 360 channels which are categorized into three groups – "Alt-Right," "Alt-Light," and "Intellectual Dark Web" – as well as a control group of popular media channels.[67] Similar to the O'Callaghan study, they identify the related channels – 2.47 million videos from 14,283 channels – and run "random walker" simulations to identify the navigation between these channels. Schmitt et al. take this approach too, utilizing two counter-messaging campaigns on YouTube and then using the API to collect data on the related videos with the aim of assessing the closeness of the content to extreme material online.[68] They then drew a randomized sample of 30% of all of the videos which were analyzed qualitatively, and a network analysis was conducted on the two datasets.

Ledwich and Zaitsev utilize both the YouTube API and a scraper to collect data on 816 channels spanning both mainstream and extreme content which are grouped into an ideological category. The study retrieves the impressions that the recommendation algorithm provides users of the channels.[69] They then test four hypotheses:

1. That recommendations influence viewers of radical content to view further radical content;
2. That the algorithm favors [mainstream] right-wing content;
3. The recommendation algorithm exposes users to more extreme content than they would otherwise seek out; and
4. The algorithm promotes a pathway from center-right or center-left to their respective extremes.

Papadamou et al. derive a set of 6,500 "incel-derived" videos that are outlinked from a range of subreddits as well as drawing a sample of 5,700 videos as a baseline for comparison. The data are accessed via YouTube's API and coded using a lexicon of incel-related words to examine videos' transcripts, metadata, and comments. To test whether the recommendation system contributes towards steering users toward incel communities, the researchers run "random walker" simulations to

66 O'Callaghan et al., "Down the (White) Rabbit Hole."

67 Ribeiro et al., "Auditing Radicalization Pathways."

68 Schmitt et al., "Counter-messages as prevention."

69 Ledwich and Zaitsev, "Algorithmic Extremism."

explore how a user could move from one video to others.[70]

Murthy identifies a seed of 11 ISIS videos that were active on YouTube in 2016.[71] He then queries the API to establish a network of the potential (a) recommended videos for the seed, (b) those recommended by the recommended videos, and (c) videos recommended by recommended videos in (b). From this he establishes a network of over 15,021 nodes and 190,087 "recommended edges." He also collects various metadata, including genre labels, view count, and comment counts, with the aim of determining whether ISIS videos were being recommended, and if so, which types of videos were recommending them. He supplements this quantitative analysis with a qualitative comparative analysis to establish whether he could identify a set of attributes that might help explain YouTube's recommendation algorithm's decision-making process.

An important point to note is that the studies discussed above which seek to find YouTube's Related Videos as a proxy for the recommendation system cannot take user personalization into account. The studies by Ribeiro et al., Ledwich and Zaitsev, and Papadamou et al. all highlight this as a limitation.[72] Instead, they provide a snapshot of what the system could recommend based on how YouTube categorizes channels. Papadamou and colleagues attempt to simulate personalization by running a random walk after watching a few incel-related videos to mimic what a user beginning to become involved in the community might do.[73] Murthy notes that he made a concerted effort not to return personalized results by using The Onion Router (TOR) browser so results would not be biased based on cookies, location, or IP address.[74]

The investigation by Whittaker et al. involved accessing Gab's API and observing the platform's different timelines, "Recent," "Popular," and "Controversial."[75] They judge the latter two to be driven by recommendation algorithms, but the former to be based on time. By comparing the timelines, they assess whether content on the "Popular" or "Controversial" timelines were more extreme than the organic flow of posts. Unlike their research on YouTube and Reddit, this does not constitute an experiment, merely a comparison between different sources of content.

Gaudette and colleagues analyze Reddit's 'upvoting' algorithm by extracting the 1000 most popular posts in the "r/The_Donald" subreddit using Reddit's API.[76] They compared these to a random sample of 1000 posts. These posts were then categorized into broad themes such as "internal threat" and "external threat". Because upvoted content is pushed to the top of users' screens, they judge whether the algorithm promotes hateful content compared to a random sample.

70 Papadamou et al., "How over is it?"

71 Murthy, "Evaluating Platform Accountability."

72 Ribeiro et al., "Auditing Radicalization Pathways"; Ledwich and Zaitsev, "Algorithmic Extremism"; Papadamou et al., "How over is it?."

73 Papadamou et al., "How over is it?"

74 Murthy, "Evaluating Platform Accountability."

75 Whittaker et al., "Recommender Systems."

76 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

4.6 Other Approaches

Other studies take a more qualitative or observational approach. Berger takes his readers through several steps, beginning with starting a new account on Twitter, following the account of al Nusra Front, and observing the site's "You Might Also Want to Follow" recommendations.[77] It would be fair to say that this piece is written more as an op-ed, yet it is included as it utilizes primary empirical evidence. As mentioned above, Waters and Postings do not seek to analyze the role of Facebook's recommendation algorithms, but in conducting their research observe the site's "Recommended Friends" algorithm.[78] Baugut and Neumann take a different approach, conducting 44 in-depth interviews with German and Austrian Islamists to explore the media diet and circumstances of their participants, which yielded findings that relate to platforms' recommendations.[79] The latter approach offers an important perspective because it puts content-sharing algorithms in the context of a wider media diet rather than focusing on them in isolation.

Coding

Account Categorization

Many of the studies deal with datasets that are too large to manually code each piece of content, instead deciding to categorize the channel from which the content comes. O'Callaghan et al. develop a set of far-right themes from the academic literature (e.g. anti-Islam; neo-Nazi) and use the retrieved text metadata to categorize channels, which were checked for reliability against Freebase, a topic annotation service provided by YouTube.[80] The Ribeiro et al. study, which collects over 300,000 videos, also categorize by channel, which are coded manually into either "Alt-Right," "Alt-Lite," or "Intellectual Dark Web."[81] This was done by collecting seed channels from the academic literature, which were independently annotated twice and disregarded if there was disagreement. For the recommendation dataset, they code the channels by having two experienced raters independently categorize the channels with 75% agreement. Where the coders disagreed, they discussed the cases until they reached consensus.

Ledwich and Zaitsev categorize channels using "soft tags" such as "Conspiracy," "Revolutionary," and "Partisan Right," as well as "hard tags" which differentiated between mainstream media sources and independent YouTubers.[82] The data is then coded by three labelers and a majority was needed to assign a categorization, with most of the values receiving an interclass correlation coefficiency that is deemed "fair" or better. Hosseinmardi et al. categorize their data into five political categories (far-left; left; center; right; and far-right),[83] drawing from the coding results outlined by Ledwich and Zaitsev

....

77   Berger, "Zero Degrees."

78   Waters and Postings, "Spiders of the Caliphate."

79   Baugut and Neumann, "Online propaganda use."

80   O'Callaghan et al., "Down the (White) Rabbit Hole."

81   Ribeiro et al., "Auditing Radicalization Pathways."

82   Ledwich and Zaitsev, "Algorithmic Extremism."

83   Hosseinmardi et al., "Evaluating."

and Ribeiro et al. above. For their study on Twitter, Huszár et al. also use a scale to categorize political partisanship; parties were determined as being "far-right" or "far-left" if their Wikipedia entries mentioned an association with far-left or far-right ideologies, or if the 2019 Chapel Hill Expert Survey indicated that the party was extreme (with a score above nine or below two).[84]

For their study on viewing behaviors on YouTube, Chen and colleagues draw from existing academic literature to identify 322 "alternative" and 290 "extremist" channels.[85] These sources included Becca Lewis' Alternative Influence report,[86] Ledwich and Zaitsev, Ribeiro et al., the Anti-Defamation League's Centre on Extremism, the Counter-Extremism Project, the Southern Poverty Law Centre, Hope Not Hate, as well channels found on the white supremacist website Stormfront.

Mixed-Method Content Classification

Some studies integrate a qualitative coding element to a portion of the content, which is then scaled up. Schmitt et al. drew a random sample of 30% of their whole dataset which was then qualitatively analyzed into categories that reflect non-extreme content (e.g. entertainment, news & politics, gaming) as well as themes such as "conspiracy theories," "hate speech," and "far-right" and "Islamist" propaganda.[87] To address the subjectivity of the coding process, the raters discuss divergent opinions until they are resolved.

Papadamou and colleagues create a lexicon of 200 incel-related words to cross-reference against YouTube videos' transcripts, metadata, and comments.[88] This is done by crawling the glossary on the incels.wiki webpage (resulting in 395 words) and then qualitatively removing general-purpose words (e.g., fuel, hole, legit, etc.). The annotators only included a word if they believed it was relevant, if it expresses hate or misogyny, or is directly associated with incel ideology. The coders worked independently and scored a Fleiss' Kappa of 0.69. They then selected a random sample of 1,000 videos that had been derived from incel subreddits and the first author manually annotated as either "incel-related" or "other." They counted the number of incel terms in the manually annotated transcript and comments, then used this as a base to automate the rest of their analysis.

Murthy begins by collecting seed ISIS videos by attribution to its media wing Al Hayat Media Center.[89] These 11 seeds led to a subset of 67 videos that recommended the seeds which are manually coded as belonging to ISIS if the group claimed the video officially and it had the group's logo. He then automates the process for the wider subset of 15,021 videos via keywords (e.g. "mujatweets," "ISIS/IS/ISIL" or an official video title). If the automated script finds a keyword, it is flagged and checked by a human researcher. Murthy then used qualitative comparative analysis to iteratively attempt to better understand the decision-making of YouTube's algorithms when recommending ISIS content.

84 Huszár et al., "Algorithmic Amplification."

85 Chen et al., "Exposure to Alternative & Extremist Content."

86 Rebecca Lewis, "Alternative Influence: Broadcasting the Reactionary Right on YouTube," (2018) https://datasociety.net/research/media-manipulation.

87 Schmitt et al., "Counter-messages as prevention."

88 Papadamou et al., "How over is it?"

89 Murthy, "Evaluating Platform Accountability."

Qualitative Content Coding

Other studies that utilized smaller datasets coded the content qualitatively in its entirety rather than the channels from which it originates. Whittaker et al. utilize the Extremist Media Index,[90] which was developed by Holbrook,[91] which categorizes content on three levels: Moderate, Fringe, and Extreme, with the latter level subdivided into four levels depending on the specificity of the threat of violence. In their study, two coders worked on a random sample of 105 pieces of content, which yielded an agreement of 80% (for a Krippendoff's alpha of 0.74). After identifying the 1000 most "upvoted" posts in r/The_Donald and a random sample of 1000 posts, Gaudette et al. also code their content line-by-line descriptively which were eventually grouped into larger qualitative categories using a thematic analysis such as "external" and "internal" threat.[92] Baugut and Neumann, who conduct interviews, utilize an interpretive qualitative content analysis.[93] They included quality checks such as "red flagging" when the authors believe a participant may have been being untruthful or had internal contradictions, as well as utilizing categories of propaganda from their bespoke radicalization model.

No Coding

Other studies did not appear to have any formal coding. Given their observational nature, the Berger[94] and Waters and Postings[95] studies do not appear to have conducted any formal coding. In their study, Wolfowicz et al. do not code online content as extremist or not, but instead use surveys to gauge participants' support for suicide bombing, as well as Twitter API data to assess their online social networks.[96]

Coding "Extremism"

As noted in the introduction, concepts such as "extremist" and many related terms (such as radicalization, terrorism, far-right, etc.) are essentially contested concepts.[97] In essence, there are no agreed-upon definitions of these words and each of them is value laden. Macdonald and Whittaker argue that the lack of conceptual clarity is particularly problematic in extremism research for three reasons: it affects the robustness of empirical research, it impedes the ability to conduct meta-reviews (such as this one), and it is difficult to articulate findings to interested parties.[98]

This conceptual ambiguity can be seen within this corpus. Although all the studies that are included in

90 Whittaker et al., "Recommender Systems."

91 Donald Holbrook, "Designing and Applying an 'Extremist Media Index,'" Perspectives on Terrorism 9, no. 5 (2015): 57–68.

92 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

93 Baugut and Neumann, "Online propaganda use."

94 Berger, "Zero Degrees."

95 Waters and Postings, "Spiders of the Caliphate."

96 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

97 Gallie, "Essentially Contested Concepts."

98 Stuart Macdonald and Joe Whittaker, "Online Radicalization: Contested Terms and Conceptual Clarity" in Online Terrorist Propaganda, Recruitment, and Radicalization, ed. John Vacca (Boca Raton: CRC Press, 2019): 33–46.

this review mention extremist content or related concepts, the approaches to coding do not all seek to classify content as "extremist," but often instead use proxy terms. For example, Ribeiro et al. frame their research as auditing "radicalization pipelines," but do so by classifying pathways among the intellectual dark web, the alt-light and alt-right.[99] This further stretches the ambiguity because each of these terms is contested: "Identifying such communities and the channels which belong to them is no easy task: the membership of channels to these communities is volatile and fuzzy, and there is disagreement between how members of these communities view themselves and how they are considered by scholars and the media."[100]

Many studies seek to classify content as "extremist" (or a related concept) by categorizing the source, such as the channel or account which produces it. This has some clear limitations. As articulated by several of the studies in this approach, the "filter bubble" or "radicalization pipeline" hypothesis is that recommender systems steer users towards more extreme content.[101] By coding channels rather than content, authors are not able to identify whether the content that users are being steered towards is actually extreme or not. Rather, to code a channel as extreme and then analyze a corpus of videos carries an assumption that every piece of content that a channel produces is equally problematic. Although the lexicon-based approach adopted by Papadamou and colleagues offers a novel solution to this problem, it has a similar issue in that it assumes that all words related to incel ideology are extreme, noting that coders were to consider a term relevant if it expressed hate, misogyny, or is directly associated with incel ideology.[102] One of the examples they offer is "Beta male," which while certainly can be used in an extremist context, has a considerably wider usage within popular culture.

An interrelated issue is classifying channels based on existing literature and online databases. Several of the studies drew from other research, such as Lewis' or the Anti-Defamation League's reports.[103] However, these original reports did not necessarily attempt to define or identify extreme content; rather, each of them analyzed or described different aspects of the contemporary far-right. Similarly, some drew from other studies in this corpus: Chen and colleagues use both Ribeiro et al. and Ledwich & Zaitsev's classifications to inform their categories of "alternative" or "extreme," even though these were not labels that the original authors used.[104] On one hand, it is good to draw from existing scholarly work to inform the research design of an empirical project, but this may come with a limitation of not being fully aligned with the original authors' intentions.

These conceptual issues are compounded when considering the legality of the content under study, which in turn affects social media platforms' obligations to remove it. Laws which proscribe against extreme content are diverse, with many countries or international organizations holding different

••

99 Ribeiro et al., "Auditing Radicalization Pathways."

100 Ribeiro et al., "Auditing Radicalization Pathways," 3.

101 For example, see Ribeiro et al., "Auditing Radicalization Pathways"; Ledwich and Zaitsev, "Algorithmic Extremism."

102 Papadamou et al., "How over is it?"

103 Lewis, "Alternative Influence"; "From Alt Right to Alt Lite: Naming the Hate," Anti-Defamation League, 2019, https://web.archive.org/web/20190422202936/ https://www.adl.org/resources/backgrounders/from-alt-right-to-alt-lite-naming-the-hate.

104 Chen et al., "Exposureto Alternative & Extremist Content."

conceptualizations as to what constitutes illegal "extreme," "terrorist," "hateful," etc. content.[105] This is particularly stark given the different transatlantic approaches to free speech, with the US providing substantially more protection.[106] Perhaps because of this complex international environment, studies in this corpus tend not to focus on whether content is illegal or not, although in some cases it can be inferred: Murthy's study is focused explicitly on ISIS videos,[107] while Berger's piece mentioned the promotion of al-Qaeda accounts,[108] both of which are designated as terror organizations in the vast majority of national lists.[109] However, when reviewing the studies that do offer either a full or partial list of channels, it is likely that the classification of "extreme" (or related concepts) falls under what the EU Counter-Terrorism Commissioner calls "legal but potentially harmful content... [which may] bring some people to embrace violent extremism."[110] This picture becomes even more complicated when considering that many studies classified by channel rather than content.

Findings

YouTube

As noted above, YouTube is by far the most popular platform under study in this corpus. Below, the results of the research on YouTube are divided into three categories: those that find the platform's content-sharing algorithms recommend extremist content; those with mixed or equivocal findings; and those that suggest that there is no "filter bubble" effect.

Positive effects

After categorizing channels, the O'Callaghan et al. study finds that there was support for the hypothesis that YouTube's algorithm recommended further far-right content, and in turn, could result in users being excluded from information that is not already in line with their ideological perspective.[111] They describe this as an "immersive ideological bubble" and argue that it places an emphasis on platforms as important political actors, whose recommendation systems are not neutral in their effects. Similarly, in their research on YouTube, Whittaker et al. find that the treatment which interacted primarily with far-right content is significantly more likely to recommend further content classified as both "Extreme" and "Fringe."[112] Conversely, the account which primarily interacted with neutral content is significantly less likely to do so, as was the baseline account. In essence, both studies show that YouTube's content-sharing algorithms are reactive to viewing extremist content and will recommend it further.

....

105 Chris Meserole and Daniel Byman, "Terrorist Definitions and Designations Lists Key Findings and Recommendations," Global Research Network on Terrorism and Technology 7 (2019).

106 Peter Neumann, "Options and Strategies for Countering Online Radicalization in the United States," Studies in Conflict and Terrorism, 36, no. 6 (2013): 431–459.

107 Murthy, "Evaluating Platform Accountability."

108 Berger, "Zero Degrees."

109 "Group Inclusion Policy," Tech Against Terrorism, (n.d.), https://www.terrorismanalytics.org/group-inclusion-policy.

110 Council of the European Union, "The Role of Algorithmic Amplification in Promoting Violent and Extremist Content and its Dissemination on Platforms and Social Media," (Brussels, 2020): 2.

111 O'Callaghan et al., "Down the (White) Rabbit Hole."

112 Whittaker et al., "Recommender Systems."

In their interview-based research on the media diet of Islamists, Baugut & Neumann find that many individuals began with a basic interest in Islam or in news media that was outside the mainstream. These individuals then followed platforms' algorithmically influenced recommendations to where they encountered radical propaganda. One incarcerated interviewee made this point:

I've never searched for it, but when you type 'Islam' on YouTube, you automatically get videos from Islamists. Or if you are looking for a preacher, then all the other [i.e., radical] preachers will also come, they have the same topics and then comes ISIS.[113]

Respondents also note that if they liked nasheed music on YouTube, then they were also recommended accompanying videos that depicted violence. Another interviewee highlights that he was propelled to act because of violent propaganda offered by YouTube, which propagated a victimhood narrative. Overall, these studies highlight these algorithms as an important and problematic factor in their participants' radicalization.

Mixed effects

Schmitt et al. find that extremist content may exist in close proximity to YouTube counter-message campaigns.[114] Both the seed accounts that they begin with (#WhatIS and ExitUSA) differed in the amount and diversity of extremist content to which they relate, which they put down to structural differences between the two campaigns. ExitUSA is mostly connected with a diverse environment of entertainment and information-related videos and only a small amount of far-right propaganda. However, they note that users can still be easily confronted with this content within two clicks via recommendations. On the other hand, some of the #WhatIS videos have a remarkable number of connections with Islamist propaganda, which they explain by the thematic overlap of specific keywords (such as "jihad") which the algorithm was assigning similarity. They note that it could be argued that extreme videos connected in the same networks as counter-messaging is positive as it gives an opportunity for the latter to reach those who are engaging with radical propaganda. However, they also note that extremists tend to be much more prolific than counter-message creators. Overall, they urge caution when crafting counter messages on social media because they may have an unintended consequence of pushing individuals closer to problematic content.

Ribeiro et al. analyze over two million YouTube recommendations that were related to their dataset of three categories: Alt-Right, Alt-Lite, and Intellectual Dark Web.[115] They find that YouTube's recommendation algorithm frequently suggests Alt-Lite and Intellectual Dark Web content, and once in these communities, it is possible to find the alt-right from recommended channels – but importantly, not from recommended videos. They qualify these findings by stating they were only able to sample a small proportion of the total recommendations and by noting that they were unable to account for personalization. Despite these limitations, they argue that their findings support the notion that there is a "radicalization pipeline" on YouTube.

113 Baugut and Neumann, "Online propaganda use," 1576.

114 Schmitt et al., "Counter-messages as prevention."

115 Ribeiro et al., "Auditing Radicalization Pathways."

In Murthy's dataset, which was collected and curated in 2016, he suggests that the chances of users finding ISIS content on YouTube are rare, but not zero.[116] He finds that when such content was recommended to users, it tended to either be from other ISIS videos, or when important metadata was shared, such as title keyword similarity – supporting the findings of Schmitt et al.[117] Importantly, he finds that the language of the video played an important role in potential recommendations; under 12% of the videos that recommended the seeds were English, while two-thirds were Arabic. This suggests that, as noted above, if research is focused on English-language content, it may skew results and miss key factors.

Papadamou and colleagues' research on incels finds that the community steadily increased over their timeframe of data collection.[118] They find that there is a small but "non-negligible amount" of incel-related videos (2.9%) within YouTube's recommendation graph recommended to users. However, their random walks suggest that if a user begins to watch incel-related videos, the algorithm recommends other incel-related content with increasing frequency. Similarly, Chen et al. also find a small but non-zero number of "extreme" or "alternative" videos recommended to its participants; over 98% of total recommendations did not lead to these types of videos.[119] However, they find that when users did watch these categories of videos, their chances of being recommended further ones increased; watching an "alternative" video led to 37% of the recommendations being further videos in this category and 2.3% being "extreme" videos. Moreover, when users watched an "extreme" video, 29% of the recommendations were to other channels in this category and 14% to an "alternative" channel. These findings are similar when considering user behavior; 98.8% of all followed recommendations were to non-extremism and non-alternative channels. However, around half of the followed recommendations from alternative or extreme videos were to further channels in their respective categories. When cross-referencing against the self-report survey, those that watched these channels were almost entirely made up of individuals that reported high levels of racial resentment, suggesting that user choices may play a bigger role than recommendation systems.

Minimal-to-no effects

The findings of one study, authored by Ledwich and Zaitsev, suggest that YouTube recommendation algorithms actively discouraged users from visiting extreme content online, which they argue refutes popular "radicalization" claims.[120] As noted above, they test four hypotheses. The first is that recommendations influence individuals to watch more content than they would otherwise have and reduce the number of alternative views. They find partial support for this: while there is a clear preference from categories towards other channels in the same category, they do not find evidence that there is a dramatic shift from extreme content to further extreme channels. Secondly, they find no support that there was a right-wing advantage within YouTube's recommendation system, instead suggesting that mainstream news is preferred. The third hypothesis is that the algorithm pushes users

116 Murthy, "Evaluating Platform Accountability."

117 Schmitt et al., "Counter-messages as prevention."

118 Papadamou et al., "How over is it?"

119 Chen et al., "Exposureto Alternative & Extremist Content."

120 Ledwich and Zaitsev, "Algorithmic Extremism."

towards more extreme content than they would otherwise have seen. They also find no support for this; their data suggests that the algorithm actually appears to restrict traffic towards extreme right-wing categories. The final hypothesis is that there is a far-right radicalization pathway that takes users from the center ground, via content that is increasingly critical of left-wing or centrist narratives, towards the extreme right. Their findings suggest that mainstream right-wing news channels such as Fox News benefit from the recommendation algorithm, but that smaller fringe YouTubers were disadvantaged, causing the authors to reject the hypothesis. Overall, Ledwich and Zaitsev reject the notion that the platform's recommendation system is a radicalization pipeline.

It is worth noting that this study has been the subject of some critique, with Ribeiro and colleagues arguing that the authors make three key mistakes.[121] Firstly, the experiment does not take personalization into account, which means that the recommendation system is more likely to suggest mainstream channels, which would not necessarily be repeated for an actual user. Secondly, the authors collected data after a major policy change (as discussed above), but they treat the findings as if they were from a previous version. Finally, the paper lacked a robust methodological explanation, such as including limitations (and adjusting them to their findings), reporting statistical significance, or providing clarity for their inter-coder reliability tests. Zaitsev responded to this critique, arguing that using an anonymous account is not a major flaw (and it was discussed in their limitations), clarifying that their claims only relate to YouTube post-policy changes, and defending their methodological transparency.[122]

The study of user preferences on YouTube by Hosseinmardi et al. also downplays the role of recommendation systems. They find that the total amount of far-right content increased over the four years under study, and those that consume it tend to show a more extreme pattern of engagement compared to those with other ideologies.[123] Importantly, they find that users self-segregate into ideological communities when watching content – i.e. an "echo chamber" effect. However, they find that the pathways that channel users towards far-right videos are diverse and much of the pathway comes from other platforms, arguing that only a fraction can be attributed to recommendations. They also find that longer sessions do not lead to more extreme content. They note that YouTube may be a source of concern given that extreme content is discoverable. However, they argue that the focus on recommendation systems is too narrow, and it is better to consider the platform as one part of a wider ecosystem. They also argue that users should be viewed as active participants, purposefully seeking out content rather than being passively recommended.

Reddit and Gab

Gaudette and colleagues' study explores whether Reddit's upvoting and downvoting algorithm facilitates "othering" discourse on the subreddit r/The_Donald. They compare the 1000 most upvoted

121   Manoel Horta Ribeiro et al., "Comments on 'Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization,'" UCL Information Security, December 29, 2019, https://sec.cs.ucl.ac.uk/posts/2019/12/youtube-radicalization-study/.

122 Anna Zaitsev, "Response to Further Critique on our Paper 'Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization,'" Medium, January 8, 2020, https://anna-zaitsev.medium.com/response-to-further-critique-on-our-paper-algorithmic-extremism-examining-youtubes-rabbit-hole-af3226896203.

123 Hosseinmardi et al., "Evaluating."

posts – which are more likely to be shown to users – to a random sample of 1000 posts. They find that the upvoted sample contained substantially more extreme discourse than the random one, which they group into two themes.[124] First was mention of the "external threat," which comprises 11.6% of the upvoted sample, made up of hateful comments about Muslims. By comparison, only 1.6% of the random sample contains this type of content. Secondly, the "internal threat" is prevalent in 13.4% of posts in the upvoted sample, which targets the left as a violent enemy that wants to hurt the West. This is compared to 5.7% in the random one. The authors note that for both groups, the random sample not only contains fewer references to the respective out-group, but the language is less extreme as well. They argue that their findings suggest that Reddit's upvoting and downvoting algorithm plays a key role in "othering" the out-groups and in turn facilitates a collective extreme identity on r/The_Donald.

Although the YouTube portion of the study by Whittaker et al. finds the recommendation system to promote extreme content, their investigations on Reddit and Gab found no such relationship.[125] Despite there being 30 pieces of content (1.4%) that were classified as "extreme" on Reddit and 416 (20%) that were judged to be "fringe," interacting with either type of content does not make the platform more likely to promote it more in future. They do find that interacting with neutral content decreased the likelihood of being recommended "fringe" content, suggesting that there may be some filtering effects. With regards to Gab, there were no significant differences between the "Latest," "Controversial," or "Popular" timelines and the amount of "Extreme" content. They do find that "Fringe" content may be prioritized in the "Popular" timeline. Overall, the authors argue that there is no evidence that either of these platforms promote extreme content via the algorithm.

Twitter

In their experimental study on new Twitter users, Wolfowicz et al. find evidence to support their hypothesis that there is an interactive relationship between filter bubbles, echo chambers, and justification of suicide bombing.[126] In each of their models, the treatment (suppressing algorithmic control on Twitter) is not found to directly affect radicalization. However, when considering the interactions between their proxy variables for echo chamber effects, they find an interactive relationship. For example, they find that for individuals in the treatment group that had more outwardly focused networks, there was a decreased likelihood of them justifying terrorism. In essence, the findings suggest that the filter bubble alone does not cause radicalization, but it could be a contributing factor, particularly when considered alongside particular types of networks. In their randomized natural experiment on Twitter, Huszár et al. find no evidence that far-right or far-left accounts are amplified on users' timelines.[127] They find that mainstream right-wing political parties and right-wing news sources are amplified because users with an algorithmically driven timeline are significantly more likely to see these sources than those with a chronological one. However, they also find that in countries where there are a substantial number of elected officials that can be described

124 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

125 Whittaker et al., "Recommender Systems."

126 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

127 Huszár et al., "Algorithmic Amplification."

as far-left or far-right (such as the Japanese Communist Party; the AfD in Germany; or VOX in Spain), the amplification is lower for these parties than for centrist parties in the same countries.

Account Recommendations

Two studies suggest that platforms may recommend extreme accounts as suggested user connections. Berger's observation of Twitter's "Who to Follow" suggestions shows that if a new user follows the now-suspended account for Syrian group al Nusra Front[128] – which was at the time associated with al-Qaeda but has since merged into Tahrir al-Sham[129] – the user was then recommended the account for the radical Ansar al-Mujahideen forum. If the user follows this account, then Twitter suggests further prominent jihadists. Importantly, the "neutral" default accounts that Twitter typically offers new users such as Justin Bieber and Lady Gaga become less frequent and are replaced by "hardcore terrorists and extremists." He also notes that the same effect can be seen if the user follows far-right accounts instead like the American Nazi Party. Waters and Postings found that "Facebook's algorithms have also actively helped connect IS supporters and build extremist networks through 'suggested friends.'"[130] They found that it was the most likely explanation for connecting two of the supporters in their sample and both of the authors discovered their own accounts recommended by jihadists after engaging with other extreme individuals. Both Berger and Waters and Postings argue that the respective platforms are inadvertently creating a network that helps to connect extremists because they desire connectivity of individuals with similar interests.

Discussion

Findings and Knowledge Gaps

Looking at the corpus of studies as a whole, the findings suggest that content-sharing algorithms may amplify extreme content towards users. Of the fifteen identified studies, only two conclude that platforms had minimal effects,[131] while one found that they actively push users away from extreme content.[132] One study had split results on different platforms – i.e. YouTube does amplify extreme content, but Reddit and Gab do not,[133] while one study found no direct effect but an interactive relationship with other variables.[134] The other ten suggested that, in various ways, recommendation systems can promote extreme content, although as noted above, this often comes with important qualifications.

Understanding the types of data that were utilized in this corpus offer an insight into the gaps in the existing literature. YouTube is by far the most researched platform within the sample, which from

....

128 Berger, "Zero Degrees."

129 Europol, "TE SAT: European Union Terrorism Situation and Trend Report," (The Hague, 2018).

130 Waters and Postings, "Spiders of the Caliphate," 78.

131 Hosseinmardi et al., "Evaluating"; Huszár et al., "Algorithmic Amplification."

132 Ledwich and Zaitsev, "Algorithmic Extremism."

133 Whittaker et al., "Recommender Systems."

134 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

one perspective is intuitive given the high level of media coverage of the platform – for example, the Christchurch killer was described by news outlets as being radicalized on YouTube.[135] However, the platform also has a researcher-friendly API which makes it more feasible to access datasets and explore the role of recommendations than other platforms.[136] In comparison, we have little empirical data as to whether Facebook algorithms promote hateful or extreme content, which is concerning given the recent testimony of whistleblower Frances Haugen, who suggests they do based on leaked internal research.[137]

More broadly, the corpus contains just five platforms, and only one can be described as "small" (Gab). Terrorists and extremists populate a range of large and small platforms, often simultaneously, to adapt to the hostile environment in which they find themselves.[138] Tech Against Terrorism find that ISIS used up to 330 different platforms,[139] while other studies (on both jihadist and far-right content) have found outlinks from Twitter to dozens of different platforms.[140] While it remains vital to understand how extremists exploit the largest platforms (as research demonstrates a complex online ecosystem in which the largest platforms still play an important role[141]), an emphasis on larger ones may lead to a knowledge gap.

The content being mostly English-language or Western-focused also may be an important gap in our understanding of potential amplification. This is likely related to the proximity or language skills of the researchers. Murthy finds that language may play an important role; Arabic videos are more likely to be recommended than English ones.[142] This may suggest that, even in an optimistic scenario in which platforms have successfully navigated the balance between removal for violative content and removing or downranking borderline content from recommendations, it could be restricted to English-language content. Platforms often trial such measures on Western audiences first; YouTube's policy of reducing borderline content was tested first in the US before being expanded outwards.[143] Murthy's findings suggest that language is an important factor and that we should not assume that policy responses have equal effects if platforms dedicate differing levels of resources to them. This is particularly important when considering the claim that Facebook's "algorithms amplified hate speech and... failed to take down inflammatory posts in the ongoing genocide against the Rohingya

135 Sam Shead, "YouTube radicalized the Christchurch shooter, New Zealand report concludes," CNBC, December 8, 2020, https://www.cnbc.com/2020/12/08/youtube-radicalized-christchurch-shooter-new-zealand-report-finds.html.

136 Whittaker et al., "Recommender Systems."

137 C-SPAN, "Facebook Whistleblower Frances Haugen testifies before Senate Commerce Committee," YouTube, October 5, 2021, https://www.youtube.com/watch?v=GOnpVQnv5Cwandab_channel=C-SPAN.

138 M. Meili Criezis, "Remaining and Expanding or Surviving and Adapting? Extremist Platform Migration and Adaptation Strategies," Global Network on Extremism and Technology, 2021, https://gnet-research.org/2021/11/12/remaining-and-expanding-or-surviving-and-adapting-extremist-platform-migration-and-adaptation-strategies/.

139 "ISIS use of smaller platforms and the DWeb to share terrorist content summary," Tech Against Terrorism, April 29, 2019, https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/.

140 Maura Conway et al., "Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts," Studies in Conflict and Terrorism 42, no. 1–2 (2018): 141–160; J. M. Berger, "The Alt-Right Twitter Census," Vox Pol (Dublin, 2018).

141 Ali Fisher, Nico Prucha, and Emily Winterbotham, "Mapping the Jihadist Information Ecosystem: Towards the Next Generation of Disruption Capability," Global Research Network on Terrorism and Technology 6 (2019).

142 Murthy, "Evaluating Platform Accountability."

143 YouTube, "Our Ongoing Work."

in Burma."[144] Importantly, the plaintiffs argue that the platform's lack of investment in local knowledge and understanding of linguistic context plays an important role. While this remains a claim, the lack of academic study in the area leaves us unable to determine the role of content-sharing algorithms.

We also lack a transparent understanding of how recommendation systems work. The studies in this corpus focus heavily on researchers externally accessing a recommendation system and then assessing its output – i.e. a "black box" approach. GIFCT's CAPPI working group note that "the limitation of such studies is that without any insight into how algorithms make recommendations, it is difficult to fully assess and understand how they may lead to different kinds of outcomes."[145] These types of approaches do not typically manipulate platforms' recommendation systems to observe their effect on users, but instead tend to manipulate real or simulated users, or merely assess the type of content that could potentially be recommended. To bridge this knowledge gap in the future, Knott and colleagues recommend empirical collaboration between external stakeholders and platforms in which internal methods can be used to study the effect of recommendation systems on users.[146] The study on Twitter by Huszár et al. is a good example of this type of collaboration,[147] but it could be expanded to attempt to understand how engaging with platforms' content – or third-party interventions – can affect behavior.[148]

An interrelated issue is that we know very little about users' choices when it comes to recommendation systems and extremist content. By their nature, most social media algorithms offer some level of user-choice (i.e., to engage with the "recommended for you" content on YouTube, or to have the personalized timeline on Twitter). One of the key debates in the broader literature on "filter bubbles" is whether they play a greater role in content selection than users' own choices.[149] This is still a largely unanswered question within the studies discussed here. Rather, most of these studies look at the "supply" of extreme content – that is to say, the environment with which users could potentially engage.[150] This leaves a causal knowledge gap as to how this may affect its audience's behaviors. The studies which analyzed user behavior offer a mixed picture. Hosseinmardi et al. and Chen et al. both offer tentative findings which suggest that users' own choices play a bigger role than recommendation systems;[151] Wolfowicz et al. suggest an interactive relationship with users' networks;[152] and Baugut and Neumann's participants suggest that YouTube's recommendation system

144 Dan Milmo, "Rohingya sue Facebook for £150bn over Myanmar genocide," The Guardian, December 6, 2021, https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence.

145 "Content-Sharing Algorithms," GIFCT, 15.

146 Alistair Knott et al., "Responsible AI for Social Media."

147 Huszár et al., "Algorithmic Amplification."

148 For example, see Erin Saltman, Farshad Kooti, and Karly Vockary, "New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis," Studies in Conflict and Terrorism, March 30, 2021.

149 For example, see Bakshy, Messing, and Adamic, "Exposure to Ideologically Diverse News,"; Ivan Dylko et al., "The Dark Side of Technology: An experimental investigation of the influence of customizability technology on online political selective exposure," Computers in Human Behavior 73 (2017): 181–190; 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 35, no. 7 (2018): 2006–2015.

150 Ines von Behr et al., "Radicalisation in the Digital Era: The use of the internet in 15 cases of terrorism and extremism" RAND Europe, 2013.

151 Hosseinmardi et al., "Evaluating"; Chen et al., "Exposure to Alternative & Extremist Content."

152 Wolfowicz, Weisburd, and Hasisi, "Examining the interactive effects."

led them towards more extreme content.[153] To better understand whether such systems actually cause harm – rather than direct users towards potentially harmful content – research must go beyond simply investigating the environment and place more emphasis on how individuals have interacted with algorithms and the choices that they have made, as well as better understanding the feedback that they receive from the platform as a result of these choices.

Contextualizing into Wider Debates

Content-sharing algorithms do not exist in a vacuum. Instead, they are just one part of the wider radical online milieu, and we should not overlook the extremist environment on the platforms the content inhabits. Put simply, for extremist materials to be recommended (and for researchers to study this using open-source methods), it must be available on social media platforms. This point is made explicit in the study by Gaudette et al., who note that much of the content does contravene the rules of r/The_Donald, but moderators did not enforce the rules, and therefore such content was allowed to remain featured in the community.[154] On the other side of the coin, Whittaker et al. argue that their null findings on Gab demonstrate the importance of the radical community of which it is part.[155] Research has shown that the platform has become a haven for problematic far-right content.[156] The authors note that there was a wide range of extreme and fringe content in their dataset, which, while not algorithmically amplified, was still easily accessible to users. Discussing YouTube, Lewis argues that although recommendation algorithms could play some role in driving extreme content, the picture is substantially more complicated;[157] she argues that content is also driven by cross-networking influencers, who can provide new and large audiences,[158] and stresses the importance of the micro-celebrity status of many influencers, which can provide an illusion of authenticity to their audiences.[159]

Despite media claims of "radicalization by algorithm," we still have a very limited understanding of what part they play in the process. The distinction between radical content and users' own choices is exemplified by Munger and Phillips, who critique the idea that YouTube's recommendation algorithm plays a prominent role. [160] They argue that this understanding is little more than an update of the now-discredited "hypodermic needle model" of mass communication. Instead, they point to the affordances that YouTube offers – notably that the primary content is video and that it is a media

153 Baugut and Neumann, "Online Propaganda use."

154 Gaudette, Scrivens, and Davies, "Upvoting Extremism."

155 Whittaker et al., "Recommender Systems."

156 Berger, "The Alt-Right Twitter Census"; Maura Conway, Ryan Scrivens, and Logan Macnair, "Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends," ICCT Policy Brief (October, 2019); Lella Nouri, Nuria Lorenzo-Dus and Amy-Louise Watkin, "Following the Whack-a-Mole: Britain First's Visual Strategy from Facebook to Gab," Global Research Network on Terrorism and Technology: Paper No. 4, 2019.

157 Rebecca Lewis, "All of YouTube, Not Just the Algorithm, is a Far-right Propaganda Machine," Medium, January 8, 2020, https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430.

158 Lewis, "Alternative Influence."

159 Rebecca Lewis, "'This Is What the News Won't Show You': YouTube Creators and the Reactionary Politics of Micro-celebrity," Television and New Media 21, no. 2 (February 2020): 201–217.

160 Kevin Munger and Joseph Phillips, "Right-Wing YouTube: A Supply and Demand Perspective," International Journal of Press/Politics 27, no. 1 (January 2022): 186–219.

company – has created a capacity to create radical alternative political cannons and communities to interpret them. These affordances create a "supply and demand" framework which highlights how YouTube has made content creation appealing for fringe political content creators, but also recognizes that regardless of these, they require an active audience to want to watch it. They note that the consumption of "white nationalist video media was not caused by the supply of this media 'radicalizing' an otherwise moderate audience. Rather the audience already existed, but they were constrained by the scope of the ideology of extant media."[161] In other words, like Lewis, they warn against over-interpreting research into the role of recommendation algorithms if it comes at the expense of understanding the audience.

There has been an increased policy concern about the role of recommendation algorithms in the radicalization process.[162] However, it is important to note that thirteen of the fifteen studies focused their research exclusively on the online domain. While it is intuitive to study an online effect in its own environment, it runs the risk of inflating the role of the Internet in cases of terrorism or extremism. Research has consistently shown that despite terrorists utilizing the Internet heavily, this does not come at the expense of offline interactions[163] and that offline social networks may play a greater role than the web.[164] Baugut and Neumann consider both domains in relation to each other. As noted above, they do suggest an important role for algorithms recommending propaganda, but also find that the online and offline domains are inseparably intertwined: "Contact with online propaganda was usually followed by personal talks with peers or preachers, and preachers used their personal talks with recruits as opportunities to show them media propaganda."[165] An over-emphasis on the online environment – including algorithms – may foster a "streetlight effect" in which research focuses on what is easily available and therefore miss the true picture of contemporary radicalization.

A final important point to consider is the potential regulation of recommendation systems. Policymakers have repeatedly signaled an intention to have greater control over how such algorithms operate. Whittaker and colleagues note that presently there is little regulation in law, and potential policy maneuvers tend to be focused on algorithmic transparency (for example, the EU's Digital Services Act or the US's Filter Bubble Transparency Act). They argue that this leaves the amplification of borderline content largely unresolved.[166] The EU Counter-Terrorism Coordinator argues that platforms should remove such content from recommendations and directly links it as being a potential conduit of radicalization.[167] However, as has been demonstrated above, this is not

161 Kevin Munger and Joseph Phillips, "A Supply and Demand Framework for YouTube Politics: Introduction to Political Media on YouTube," Penn State Political Science (October, 2019): 12.

162 HM Government, "Online Harms White Paper," (London: The Stationary Office, 2019); Council of the European Union, "The Role of Algorithmic Amplification."

163 von Behr et al., "Radicalisation in the Digital Era"; Paul Gill et al., "Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns, and Processes," Criminology and Public Policy 16, no. 1) (2017): 99–117; Joe Whittaker, "The Online Behaviors of Islamic State Terrorists in the United States," Criminology and Public Policy 20, no. 1 (2021): 177–203; Chamin Herath and Joe Whittaker, "Online Radicalisation: Moving Beyond a Simple Dichotomy," Terrorism and Political Violence (November 22, 2021).

164 Sean C. Reynolds and Mohammed M. Hafez, "Social Network Analysis of German Foreign Fighters in Syria and Iraq," Terrorism and Political Violence 31, no. 4 (April, 2019): 661–686.

165 Baugut and Neumann, "Online Propaganda use," 1585.

166 Whittaker et al., "Recommender Systems."

167 Council of the European Union, "The Role of Algorithmic Amplification."

a link that is supported by evidence. Moreover, identifying such content is particularly problematic and undoubtedly leaves sizable grey areas. Tech Against Terrorism argue against regulatory policies of removing legal yet harmful content from recommendations, suggesting that it has negative implications for freedom of speech, the rule of law, and raises serious concerns over extra-legal norm-setting.[168]

Recommendations

1. **Broader scope:** This review demonstrates that most of the body of knowledge is drawn from research that focuses on English-language content, in Western countries, with a focus on the far-right. However, there may be good reason to believe that platforms are more finely attuned to this context at the expense of other locations, languages, and ideologies.

2. **More internal research:** To provide the clearest picture of the phenomenon, stakeholders should foster collaboration between those with access to recommendation algorithms (i.e. social media platforms) and researchers. Policymakers can aid this process by making data-sharing exemptions explicit within data protection regulations.

3. **Audit platforms' responses:** This review highlighted how platforms changed their policies on content recommendations over the last decade. However, academic research has yet to establish whether these responses are effective. For example, are authoritative voices raised into the recommendations of borderline content, or can groups tied to offline violence still be found in recommendations?

4. **Account for personalization and track user behavior:** Existing research mostly focuses on the "supply" of content that could potentially be recommended to users. Future research should utilize experimental designs which account for platforms' personalization rather than gathering potential recommendations. Similarly, future research should focus on "demand" by using methodological designs which track user behavior (i.e. which can show not only whether content is recommended but whether users actually follow these recommendations).

5. **Code content rather than accounts:** Many of the studies in this corpus classified accounts/channels as "extremist" (or a related term). This carries an implicit assumption that all content from the account is equally problematic. However, this is clearly not the case. If future studies' research questions relate to whether extreme content is being amplified, then it should be content that is coded.

6. **Researcher Transparency:** Researchers should offer clear explanations of the methodological decisions that are taken. This includes giving working definitions for coding, particularly when referring to essentially contested concepts, conducting and detailing inter-rater reliability tests, and providing information on how data are coded (including making datasets available if possible).

7. **Platform Transparency:** Platforms should offer an explainable rationale for why users have been recommended content. This should be available for researchers to explore empirically in future studies.

168 "Content Personalization and the Online Dissemination of Terrorist and Violent Extremist Content," Tech Against Terrorism, 2021, https://www.techagainstterrorism.org/wp-content/upload s/2021/02/TAT-Position-Paper-content-personalisation-and-online-dissemination-of-terrorist-content1.pdf.
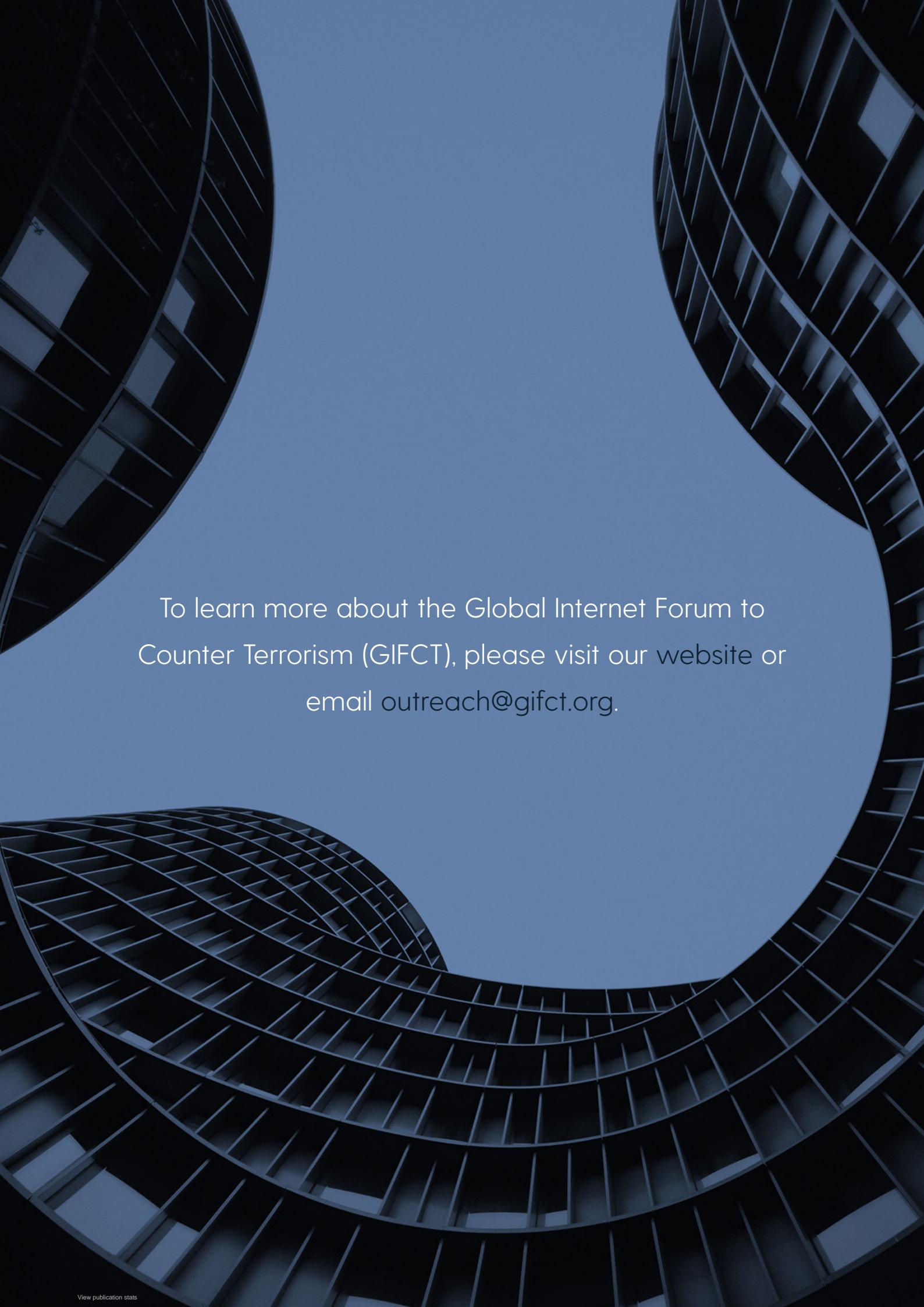
Appendix

Study	Platform	Ideology	Language	Methods	Findings
Berger (2013)	Twitter	Jihadist	Arabic	Exploration of Twitter's recommendation system. Creates new account and follows jihadist accounts.	"Who to follow" recommends a number of prominent jihadist accounts.
O'Callaghan et al. (2015)	YouTube	Far-right	English; German	Access API to draw Related Videos. Use text metadata to categorize channels, which were checked against Freebase.	Recommends further far-right content that could result in "immersive ideological bubble."
Schmitt et al. (2018)	YouTube	Jihadist; Far-right	English; German	Access API to collect Related Videos for two counter-messaging campaigns. Qualitatively analyzed and categorized 30% of dataset.	Extremist content within related videos. High crossover with anti-jihadist campaign (possible due to keyword similarity).
Waters & Postings (2018)	Facebook	Jihadist	Multiple	Social network analysis.	At least two ISIS supporters likely recommended as friends. Authors were also recommended IS-supporting accounts.
Ledwich & Zaitsev (2019)	YouTube	Far-right	English**	Access API and use scraper to collect data on seed channels. Code into categories based on ideology and mainstream vs independent.	YouTube actively discourages users from extreme content. No evidence to suggest movement towards more extreme categories.
Ribeiro et al. (2019)	YouTube	Far-right	English**	Audit seed channels which have been categorized into ideological groups. Access API to identify Related Videos and simulate navigation between channels.	YouTube recommends "Alt-Lite" and "Intellectual Dark Web" content, and once in these communities it is possible to find "Alt-Right" content, but not from recommendations. Suggest that findings support the notion of a "radicalization pipeline".
Gaudette et al. (2020)	Reddit	Far-right	English	Compare 1000 most "upvoted" posts in "r/The_Donald" against random sample.	Most upvoted sample substantially more extreme than random sample.
Baugut & Neumann (2020)	n/a*	Jihadist	German	44 interviews to explore media diet.	Individuals said that platform recommendations took them from basic knowledge to radical propaganda.
Hossein-mardi et al. (2020)	YouTube	Far-right	English	Representative sample of web users' browser history over 4 years. Channels coded according to political ideology.	Pathways towards far-right content is diverse and only a fraction can be attributed to recommendations. No trend towards more extreme content over longer sessions. Suggest user preference plays a bigger role.
Wolfowicz et al. (2021)	Twitter	Jihadist	Arabic	Recruited 96 non-Twitter users. Treatment group suppresses algorithm, control group accepts all automated suggestions. Ask participants how they feel about suicide bombing.	Interaction effect between recommendations and network effects (i.e. filter bubble and echo chamber are complementary).

Whittaker et al. (2021)	YouTube; Reddit; Gab	Far-right and male supremacist	English	YouTube/Reddit: Create identical accounts, use bot to log in engage with content. Access recommendations via API. Gab: Access data via API to compare "Recent," "Popular," and "Controversial" timelines. All: Code data according to Extremist Media Index (Holbrook 2015).	YouTube: Extreme and Fringe content more likely to be recommended and to be ranked higher. Reddit/Gab: Extreme material not promoted via recommendations.
Papadamou et al. (2021)	YouTube	Incel	English**	Compare 6.5k incel videos against 5.7k random videos. Build lexicon of 200 incel-related words to code videos as incel-related from transcript. Access via YouTube API Conduct Random Walker simulation.	Small chance that users will be recommended incel videos by system. If user watches incel-related videos, algorithm recommends other incel-related videos with increasing frequency.
Chen et al. (2021)	YouTube	Far-right	English	Link US nationally-representative survey to YouTube viewing behaviors. Identify "alternative" and "extreme" accounts via literature.	YouTube recommendations can expose users to potentially harmful content. However, vast majority of exposure is to individuals with self-reported racial resentment.
Murthy (2021)	YouTube	Jihadist	Arabic; English; French; Mandarin	Detect 11 ISIS videos as seeds. Access API to establish network that represents 1) recommended videos, 2) recommendations of recommendations, and 3) the recommendations of (2). Use qualitative comparative analysis to assess which features likely influence algorithmic decision-making.	Chance of finding ISIS content accidentally is rare, but non-zero. Usually happened when there was similar metadata. When ISIS content was recommended, it tended to be from other ISIS videos (particularly those that were not English language). Radical keywords seem to be important in recommending ISIS.
Huszár et al. (2022)	Twitter	Far-right and Far-left	English; Japanese; French; Spanish; German; Turkish	Identified 3634 legislators from seven countries. 1% of Twitter users were excluded from timeline personalization (Control group). Compare the reach of legislators from control group to treatment group who do have personalized timeline. Sought to assess whether XR/XL legislators have greater reach in personalized timelines.	No evidence to support that far-right or far-left groups are amplified more than moderate ones.

*Data derived from interviews

** Not explicitly stated but examples or keywords are entirely or primarily English-language

To learn more about the Global Internet Forum to Counter Terrorism (GIFCT), please visit our website or email outreach@gifct.org.