INTERNET POLICY REVIEW
Journal on internet regulation

Volume 10 **Issue** 2

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Recommender systems and the amplification of extremist content

Joe Whittaker *Swansea University* j.j.whittaker@swansea.ac.uk
Seán Looney *Swansea University*
Alastair Reed *Swansea University*
Fabio Votta *University of Amsterdam*

DOI: https://doi.org/10.14763/2021.2.1565

Published: 30 June 2021
Received: 11 May 2020 **Accepted:** 4 December 2020

Funding: This research was supported by the Global Research Network on Terrorism and Technology.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1565

Keywords: Filter bubble, Online radicalisation, Algorithms, Extremism, Regulation

Abstract: Policymakers have recently expressed concerns over the role of recommendation algorithms and their role in forming "filter bubbles". This is a particularly prescient concern in the context of extremist content online; these algorithms may promote extremist content at the expense of more moderate voices. In this article, we make two contributions to this debate. Firstly, we provide a novel empirical analysis of three platforms' recommendation systems when interacting with far-right content. We find that one platform—YouTube—does amplify extreme and fringe content, while two—Reddit and Gab—do not. Secondly, we contextualise these findings into the regulatory debate. There are currently few policy instruments for dealing with algorithmic amplification, and those that do exist largely focus on transparency. We argue that policymakers have yet to fully understand the problems inherent in "de-amplifying" legal, borderline content and argue that a co-regulatory approach may offer a route towards tackling many of these challenges.

Introduction

Recent years have seen a substantial increase of concern by policymakers towards the impact and role of personalisation algorithms on social media users. A key concern is that users are shown more content with which they agree at the expense of cross-cutting viewpoints, creating a false sense of reality and potentially damaging civil discourse (Vīķe-Freiberga et al., 2013). While human beings have always tended to gravitate towards opinions and individuals that align with their own beliefs at the expense of others (Sunstein, 2002), so called "filter bubbles" are singled out as being particularly problematic because they are proposed to artificially exacerbate this confirmation bias without a user's knowledge (Pariser, 2011).

The filter bubble debate is complex and often ill-defined; encompassing research into timelines, feeds, and news aggregators (Bruns, 2019), often trying to establish the importance of these algorithms versus a user's own personal choice (see e.g., Bakshy, Messing, and Adamic, 2015; Dylko et al., 2017). We lay our contribution on one specific aspect of the debate—the role of social media platforms' recommendation systems and interaction with far-right extremist content. Policymakers have articulated concern that these algorithms may be amplifying problematic content to users which may exacerbate the process of radicalisation (HM Government, 2019; Council of the European Union, 2020). This has also become a concern in popular news media, spawning articles which highlight the importance of recommender systems in *The Making of a YouTube Radical* (Roose, 2019), or referring to the platform as *The Great Radicalizer* (Tufekci, 2018) or as having radicalised Brazil and caused Jair Bolsonaro's election victory (Fisher & Taub, 2019). Research into this phenomenon (discussed in more detail below) does seem to support the notion that recommendation systems can amplify extreme content, yet the empirical research tends to focus on a single platform—YouTube—or is observational or anecdotal in nature.

We contribute to this debate in two ways: firstly, we conduct an empirical analysis of interactions of recommendation systems and far-right content on three platforms—YouTube, Reddit, and Gab. This analysis provides a novel contribution by being the first study to account for personalisation in an experimental setting, which has been noted as a limitation by previous empirical research (Ledwich and Zaitsev, 2019; Ribeiro et al., 2019). We find that one platform—YouTube—does promote extreme content after interacting with far-right materials, but the other two do not. Secondly, we contextualise these findings into the policy debate, surveying the existing regulatory instruments, highlighting the challenges that are faced, before arguing that a co-regulatory approach may offer the ability to overcome these

challenges while providing safeguards and respecting democratic norms.

Filter bubbles

Originally articulated by Eli Pariser, the concept *filter bubbles* posits that personalisation algorithms may act as "autopropaganda" by invisibly controlling what web users do and do not see, promoting ideas that users are already in agreement with and, in doing so, dramatically amplifying confirmation bias (Pariser, 2011). He claims that the era of personalisation began in 2009 when Google announced that they would begin to filter search results based on previous interactions which creates a personalised universe of information for each user, which may contain increasingly less diverse viewpoints and therefore increase polarisation within society. Since Pariser's original contribution, online platforms have further developed and emphasised their personalisation algorithms beyond "organic" interactions. De Vito notes that Facebook's News Feed is 'a constantly updated, personalized machine learning model, which changes and updates outputs based on your behaviour…Facebook's formula, to the extent that it actually exists, changes every day' (DeVito, 2016, p. 16), while YouTube's marketing director told a UK House of Commons select committee that around 70% of content watched on the platform was derived from recommendations rather than users' organic searches (Azeez, 2019). Many scholars have highlighted concerns over such algorithms, such as them carrying the biases of their human designers (Bozdog, 2013); conflating the distinction between user satisfaction and retention (Seaver, 2018), which may blind users to important social events (Napoli, 2015); people not being aware of how they restrict information (Eslami et al., 2015; Bucher, 2017); and even their creators not fully understanding how they operate (Napoli, 2014; DeVito, 2016). In the context of extremism, each of these concerns may cause alarm as they posit a situation in which users may become insulated with confirming information in an opaque system.

In turn, a widespread concern has grown in policy circles. The EU Group on Media Freedom and Pluralism suggests that it may have adverse effects: 'Increasing filtering mechanisms makes it more likely for people to only get news on subjects they are interested in, and with the perspective they identify with…Such developments undoubtedly have a potentially negative impact on democracy' (Vīķe-Freiberga *et al*., 2013, p. 27). More recently, the filter bubble effect has been blamed as a critical enabler of perceived failures of democracy such as the populist elections of Donald Trump, Jair Bolsonaro, and the Brexit referendum (Bruns, 2019). In the US, lawmakers have been considering legislative options to regulate

social media algorithms. The proposed bipartisan "Filter Bubble Transparency Act" which, at time of writing is awaiting passage through the US Senate, claims to 'make it easier for internet platform users to understand the potential manipulation that exists with secret algorithms' (Filter Bubble Transparency Act, 2019, S, LYN19613).

Importantly to the aims of this paper, the role of personalisation algorithms promoting extreme content has been highlighted by policymakers as a prescient concern. After the Christchurch terror attack in 2019, New Zealand Prime Minister Jacinda Arden and French President Emmanuel Macron brought together several heads of state and tech companies to propose the Christchurch Call. The Call committed governments to, amongst other things, 'review the operation of algorithms and other processes that may drive users towards and/or amplify terrorist and violent extremist content' (Christchurch Call, 2019). The Call was signed by the European Commission, the Council of Europe, and 49 nation states. The UK government has also signalled personalisation algorithms as problematic in its *Online Harms White Paper*, noting that they can lead to echo chambers and filter bubbles which can skew users towards extreme and unreliable content (HM Government, 2019). Similarly, the UK's Commission for Countering Extremism called on tech companies to 'ensure that their technologies have a built-in commitment to equality, and that their algorithms and systems do not give extremists the advantage from the start by feeding existing biases' (Commission for Countering Extremism, 2019, p. 86). In 2020, the EU Counter-Terrorism Coordinator released a report which argued that online platforms are a conduit for polarisation and radicalisation because their recommendation systems promote content linked to strong negative emotions, including extreme content (Council of the European Union, 2020).

Many scholars have criticised *filter bubbles* as being a problematic concept. Bruns notes that Pariser's original formulation as being founded in anecdotes, for which there is scant empirical evidence. He argues that the disconnect between the public understanding of the concept and the scientific evidence has all the hallmarks of a moral panic which distract from more important matters, such as changing communications landscapes and increasing polarisation (Bruns, 2019). Munger & Phillips also criticise the prominent theory of YouTube radicalisation via algorithm, which is often purported in the media. They argue that this argument is tantamount to the now-discredited "Hypodermic Needle" model of mass communications and instead posit a "supply" and "demand" framework which emphasises both the affordances that the platform offers as well as a greater focus on the audience (Munger and Phillips, 2019).

Despite concerns by academics and policymakers, there is limited empirical evidence as to the extent (or harmfulness) of filter bubbles (Zuiderveen Borgesius *et al*., 2016); research tends to suggest that social media users have a more diverse media diet than non-users (Bruns, 2019). Studies have shown that personalisation algorithms do filter towards an ideological position and can increase political polarisation (Bakshy, Messing, and Adamic, 2015 [1]; Dylko *et al*., 2018) but that they play a smaller role than users' own choices. Several studies have also studied the role of news recommendation systems, finding that the concerns over personalised recommendations are overstated (Haim, Graefe, and Brosius, 2018), do not reduce the diversity of content to users (Möller *et al*., 2018), and are more likely to be driven by factors such as time or date than past behaviours (Courtois, Slechten, and Coenen, 2018).

A key part of most of the critiques of the filter bubble as a concept is that it lacks clarity. There is little distinction made between the similar, yet different, concepts of filter bubbles and echo chambers (Zuiderveen Borgesius *et al*., 2016; Bruns, 2019). This is problematic because it hinders ability to do robust research, studies may offer different findings because they employ radically different definitions of these concepts (Bruns, 2019). It is worthwhile to consider the distinction drawn by Zuiderveen Borgesius and colleagues, who distinguish between two types of personalisation: firstly, self-selected, in which the user chooses to encounter like-minded views and opinions (i.e., an "echo chamber") and secondly, pre-selected, which is driven by platforms without the user's deliberate choice, input, knowledge or consent—i.e., a "filter bubble" (Zuiderveen Borgesius *et al*., 2016). This conceptual confusion also extends to discussions of involvement in terrorism and extremism, Whittaker (2020) argues that studies have frequently posited a causative relationship between online echo chambers and radicalisation—with little empirical evidence—*and* they are rarely clear as to whether they refer to users' own choices or the effects of algorithms.

This conceptual distinction is important for the present study because social media recommendation systems do not fit easily into either the self-selected or pre-selected categories. In all three of the platforms in consideration for the present study (YouTube, Reddit, and Gab), content is pre-selected for users, with which they then have the option to engage. This is different, for example, to Facebook's News Feed in which users do not have this option. Given this ambiguity, it is worthwhile to be clear about the phenomena that are under consideration—recommendation

1. It is worthwhile to note that this study was commissioned and undertaken by Facebook employees and studied the platform in question.

systems—which are defined by Ricci, Rokach and Shapira (2011) as software tools and techniques providing users with suggestions for items a user may wish to utilise. They expand upon this by stating 'the suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read' (Ricci, Rokach and Shapira, 2011, p. 1). A user *can* navigate these platforms without utilising these systems—even if they happen to make up most of the traffic. Moreover, we define content amplification as the promotion of certain types of content—in this case far right extreme content—at the expense of more moderate viewpoints.

Extremist content and recommendation systems

There are a handful of existing studies which seek to explore the relationship between recommendation algorithms and extremist content, with a predominance towards YouTube. O'Callaghan *et al.* (2015) conducted an analysis of far-right content, finding that more extreme content can be offered to users and they may find themselves in an immersive ideological bubble in which they can be excluded from content which does not fit their belief system. Ribeiro *et al.* (2019) conducted an analysis of two million recommendations that were related to three categories: Alt-right, [2] Alt-Lite, [3] and Intellectual Dark Web. [4] 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. This, they argue, supports the notion that YouTube has a "radicalisation pipeline". Schmitt *et al.* (2018) study YouTube recommendations in the context of counter-messages designed to dissuade individuals from extremism. They utilise two seed campaigns (#WhatIS and ExitUSA), finding that the universe of *related videos* had a high crossover with radical propaganda, particularly the #WhatIS campaign which had several key words with thematic overlaps with Islamist propaganda (for example: "jihad"). All three studies' findings suggest that extremist content may be amplified via YouTube's recommendation algorithm.

Conversely, Ledwich and Zaitsev (2019) find that YouTube recommendation algo-

2. In their paper, they use the Anti-Defamation League's description of the Alt-Right as: "Aloose segment of the white supremacist movement consisting of individuals who reject mainstream conservatism in favor of politics that embrace racist, anti-Semitic and white supremacist ideology" (Ribeiro et al., 2019, p. 2).
3. They argue that Alt-lite was created to demarcate individuals and content that engage in civil nationalism, but deny a link to white supremacy.
4. A group of contrarian academics and podcast hosts who discuss and debate a range of social issues such as abortion, LGBT issues, identity politics, and religion.

rithms actively discourage users from visiting extreme content online, which they claim refutes popular "radicalisation" claims. They develop a coding system for channels based on ideological categories (e.g., "Conspiracy", "Revolutionary", "Partisan Right") as well as whether the channel was part of the mainstream news or independent YouTubers. They find no evidence of migration to extreme right channels—their data suggest that the recommendation algorithm appears to restrict traffic towards these categories and that users are instead directed towards mainstream channels. Importantly, all four of these studies generate data by leveraging YouTube's application programming interface (APIs) to retrieve data from respective platforms to analyse recommendations of extremist content. The data collection does not mimic the user-platform relationship that typically has users interact with content, includes repeated exposure to recommendations and continued interactions with content from which an algorithm can learn and tailor content accordingly. As such, these studies do not consider personalisation which is at the core of the filter bubble hypothesis. This is acknowledged by both Ribeiro *et al.* (2019) and Ledwich and Zaitsev (2019) as limitations.

Other studies have taken qualitative or observational approaches. Gaudette *et al.* (2020) studied Reddit's upvoting and downvoting algorithm on the subreddit r/The_Donald by taking a sample of the 1,000 most upvoted posts and comparing them to a random sample, finding that the upvoted sample contained extreme discourse which facilitated "othering" towards two outgroups—Muslims and the left. The authors argue that the upvoting algorithm plays a key role in facilitating an extreme collective identity on the subreddit. Baugut and Neumann (2020) conduct interviews with 44 radical Islamists on their media diet, finding that many individuals began with only a basic interest in ideology but followed platforms' recommendations when they were shown radical propaganda, which propelled them to engage in violence. Both Berger (2013) and Waters and Postings (2018) observe that Twitter and Facebook respectively suggest radical jihadist accounts for users to follow after an individual begins to engage with extreme content, arguing that the platforms inadvertently create a network which helps to connect extremists.

Looking at the existing literature, several inferences can be made. Firstly, little is known about the relationship between recommendation systems and the promotion of extremist content; of the few studies that exist, a majority do suggest that these algorithms can promote extreme content, but it tends to focus on one platform—YouTube—and mostly analyses the potential interaction between users and content or relies on collecting potential recommendations from platform APIs, rather than actual interactions based on personalisation. The focus on a single

platform is significant because research may be driven by convenience to researchers due to YouTube's open API rather than following the trail of extreme content. Secondly, when looking to the previous section and research into filter bubbles more broadly, despite theoretical apprehensions which can be alarmingly applied to the promotion of extremist content, the evidence base for this concern is limited, most studies tend to play down the effect, either suggesting that it does not exist or that users' own choices play a bigger role in their decision-making.

Methodology

This study aims to empirically analyse whether social media recommendation systems can promote far-right extremist content on three platforms —YouTube, Reddit, and Gab. The Far-right was chosen to be the most appropriate ideology because it remains accessible on the internet because platforms have not been able to utilise the same methods of de-platforming as they used on jihadist content (Conway, 2020). Each of the platforms in question have been noted as hosting extreme far-right material, for example: YouTube (O'Callaghan *et al.*, 2015; Lewis, 2018; Ottoni *et al.*, 2018; Munger and Phillips, 2020; Van Der Vegt *et al.*, 2020), Reddit (Conway, 2016; Copland, 2020; Gaudette *et al.*, 2020), and Gab (Berger, 2018b; Conway, Scrivens, and Macnair, 2019; Nouri, Lorenzo-Dus, and Watkin, 2019).

Our methodology adds an important contribution to the existing literature. Rather than utilising the results from an API to generate data on potential recommendations, which is common in the existing literature (O'Callaghan et al., 2015; Ledwich and Zaitsev, 2019; Ribeiro et al., 2019), we engage with social media recommendation systems via automated user accounts (bots) to observe content adjustments (recommendations). Retrieving potential recommendations via the API does not mimic the users' relationship to the platform because the algorithm has nothing to learn about an individual user and no personalisation takes place. On the other hand, our methodology utilises automated agents which create behavioural data that the algorithm uses to personalise recommendations. Therefore, we effectively recreate the conditions in which a real user finds itself when using a platform with a personalised recommendation system. A similar research design was conducted by Haim, Graefe and Brosius (2018), who created personalised accounts on Google News in order to study the political bias of news recommendations. However, there are no studies that utilise this methodology to assess the amplification of extremist content. Given the divergent platform architecture, we utilise three different designs, explained below.

YouTube

We investigate whether extreme content was promoted after applying specific treatments. To do this, we created three identical accounts subscribed to the same 20 YouTube channels: 10 far-right channels that were identified in the academic literature [5] and 10 apolitical content producers (for example, sport or weather). We subjected these accounts to three different treatments:

1. Acting predominantly with far-right channels—the extreme interaction account (EIA);
2. Acting predominantly with apolitical channels—the neutral interaction account (NIA); and
3. Doing nothing at all—the baseline account (BA).

Data were collected by visiting the YouTube homepage twice per day. Using the recommendation data, we proceeded to construct two variables of interest —the *share* and the *rank* of extreme, fringe, and moderate content. For the share of content, we divide the respective content of a specific category by the total number of content pieces by recommendation set. To determine rank, content that appears on top left is ranked as "1" so that for each content piece to the right and below the rank continuously increases. YouTube offers 18 recommended videos, so the data have a ranking of 1-18. This serves as a measurement of algorithm prioritisation as we can assume that content that is more easily accessible (e.g. more visible) will be consumed and viewed more.

For the first week, the three accounts did not interact with any content and just visited the homepage twice a day to collect data. This serves as a baseline to be compared against the treatments. After a week, the three different treatments were applied. 20 videos were chosen (one from each channel), and all the accounts watched one to kickstart the recommendation algorithm. Each time an account visited the YouTube frontpage, ten videos were randomly chosen from the recommendation tab. For the EIA, seven videos were chosen from far-right channels and three from neutral. The NIA watched seven neutral channels and three far-right. If this operation was not possible to perform because there were not enough videos from neutral or extreme channels present, videos were watched twice until the quota was met. If no video appeared from any of the initial 20 channels in any given session, the account would watch a video from the initial 20 videos that were used to kickstart the algorithm.

5. The authors follow the policy of Vox-Pol and J M Berger by not identifying the names of accounts in this research, both for reasons of potentially increasing exposure and privacy. See, Berger (2018).

Quasi-Poisson models were used to estimate rate ratios and expected frequency counts to test whether extremist or fringe content was more or less prevalent after treatments were applied (Agresti, 2013). To test for rank differences in content, Wilcoxon rank sum tests were chosen, a non-parametric alternative to the unpaired two-samples t-test, which was chosen due to non-normality of the rank distribution (Kraska-Miller, 2013).

Reddit

The design for the Reddit experiment is almost identical to that of YouTube. We created three identical accounts and followed the same selection of far-right (including male supremacist) and apolitical Subreddits. We left these accounts dormant for one week to collect baseline data, before conducting three treatments:

1. Extremist Interaction Account (EIA), which acted predominantly with far-right content
2. Neutral Interaction Account (NIA), which acted predominantly with neutral content
3. Baseline Account (BA), which did not interact at all.

As with YouTube, the two variables of interest were the *share* and *rank* of extreme and fringe content, which was decided based on where content appeared on Reddit's "Best" timeline. Reddit offers 25 posts from top-to-bottom: giving the top result a rank of 1 through to the bottom of 25. The same procedure was followed as highlighted above; we collected data twice per day by logging in and viewing the recommended post, with the EIA interacting with seven posts from far-right subreddits and three apolitical ones, and the NIA interacting with seven apolitical and three far-right. Again, Quasi-Poisson models were used to estimate rate ratios and expected frequency counts, and Wilcoxon rank sum tests were utilised to test differences in rank.

Gab

Gab's architecture is fundamentally different—and substantially more basic—requiring a more simplistic approach. One function that Gab offers is the ability to choose between three different types of news feed: "Popular", "Controversial", and "Latest". Although not made explicit by the platform, we judge the first two to be algorithmically driven by non-chronological factors, possibly related to Gab's upvote and downvote system. However, "Latest" by definition, is based either entirely or primarily on the most recent posts, which offers the ability to analyse how algorithmically recommended posts compare against a timeline influenced primarily by

recency. We collected data from each of the three timeline options for three of Gab's topics: News, Politics, and Humour, creating nine different investigations in total. We then assessed how much extreme content appeared in each timeline.

The experiments on YouTube and Reddit are relatively similar in aims and scope, while Gab's investigation diverges. Therefore, the research questions are as follows:

- *RQ1: Does the amount of extreme content increase after applying treatments? (YouTube & Reddit)*
- *RQ2: Is extreme content better ranked by the algorithm after applying treatments? (YouTube & Reddit)*
- *RQ3: Do Gab's different timelines promote extreme content? (Gab)*

Data were collected over a two-week period in January/February 2019. For YouTube and Reddit, the bots logged in twice per day, which created 28 different sessions, for a total of 1,443 videos (of which 949 were unique), and 2,100 posts on Reddit (of which 834 were unique). Unfortunately, during the data collection period, Gab experienced several technical issues resulting in disruptions to the site, meaning that the authors were only able to log in for five sessions. This still resulted in 1,271 posts being collected, of which 746 were unique, which we deemed adequate for an exploratory investigation.

Coding

Two members of the team coded the data using the *Extremist Media Index* (EMI), which was developed by Donald Holbrook and consists of three levels: Moderate, Fringe, and Extreme (Holbrook, 2015, 2017b, 2017a). For an item to be categorised as Extreme, it must legitimise or glorify the use of violence or involve stark dehumanisation that renders an audience sub-human. For Holbrook's research, this category also includes four sub-levels which relate to the specificity and immediacy of the violence, but for this study, the sample sizes were too small to produce reliable results. To be deemed Fringe, content had to be radical, but without justifications of violence. Anger or blame might also be expressed towards an out-group and may include profanity laden nicknames that go beyond political discourse (e.g., "libtards" or "feminazis"), or historical revisionism. All other content was deemed as Moderate, which can include references to specific out groups if it is deemed a part of acceptable political discourse.

For inter-rater reliability, the two categorisers coded a random sample of 35 pieces

of content from each of the three platforms (105 in total). The two raters agreed in 80.76% of cases: 74.3% on YouTube, 85.7% on Reddit, and 81.8% on Gab, yielding a Krippendoff's alpha value of 0.74 (YouTube = 0.77, Reddit = 0.72, Gab = 0.73). These values are deemed acceptable to draw tentative conclusions from the data. The coders then categorised the remaining content on each platform. Only the original post/video was taken into account (i.e., not comments underneath or outward links). In YouTube, many of the videos were multiple hours long, therefore raters had to make their decision based on the first five minutes.

Two thirds of the collected data on YouTube was rated as Moderate (n=949), while 28% was Fringe (409), and 6% was Extreme (85). On Reddit, almost four-fifths of the content was classified as Moderate (1,654) while 20% was Fringe (416) and less than 2% was Extreme (30). On Gab, 64% of posts were deemed to be Moderate (810), with 29% Fringe (366) and 7% was Extreme (95).

Results

RQ1: Does the amount of extreme content increase after applying treatments?

On YouTube, we found that the account that predominantly interacted with far-right materials (the EIA) was twice as likely to be shown Extreme content, and 1.39 times more likely to be recommended Fringe content. Conversely, the NIA and BA were 2.96 and 3.23 times less likely to be shown Extreme content. These findings suggest that when users interact with far-right content on YouTube, it is further amplified to them in the future.

On the other hand, Reddit's recommendation algorithm does not seem to promote Extreme content with the EIA; none of the models show statistically significant effects, suggesting that interacting with far-right content does not increase the likelihood that a user is recommended further extreme content.

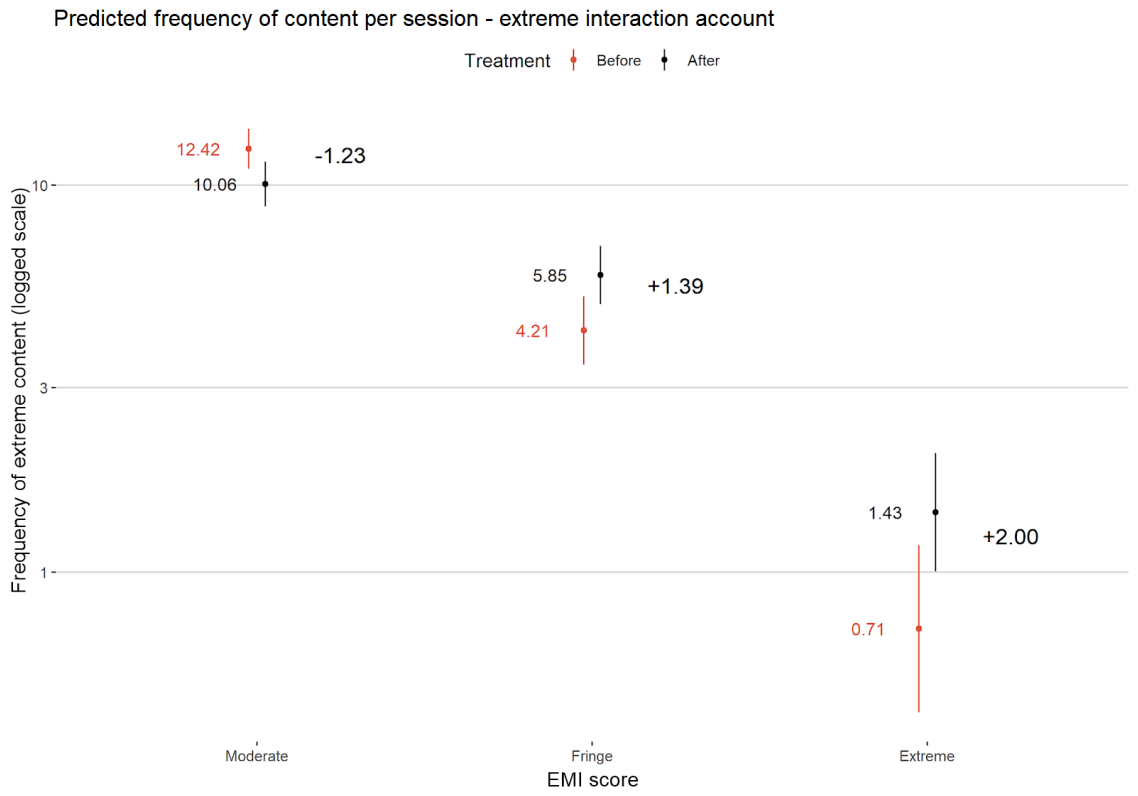

FIGURE 1: Predicted frequency of content per session (YouTube) for the EIA condition from Quasi-Poisson model. 95% confidence interval shown.

RQ2: Is extreme content better ranked by the algorithm after applying treatments?

As with RQ1, we found that YouTube prioritises Extreme content; it ranked such content significantly higher than Moderate. In the EIA the median rank for the former was 5, while the latter was 10. There was no significant difference between the Fringe and Extreme or the Fringe and Moderate content. There was also no significant difference between the EMI categories in the NIA or BA.

The results for RQ2 on Reddit also mirror those of RQ1; we found no statistically significant differences between any of the variables in the EIA. We did observe that the NIA does decrease the average rank of fringe content on the platform, which does point to a minor filtering effect.

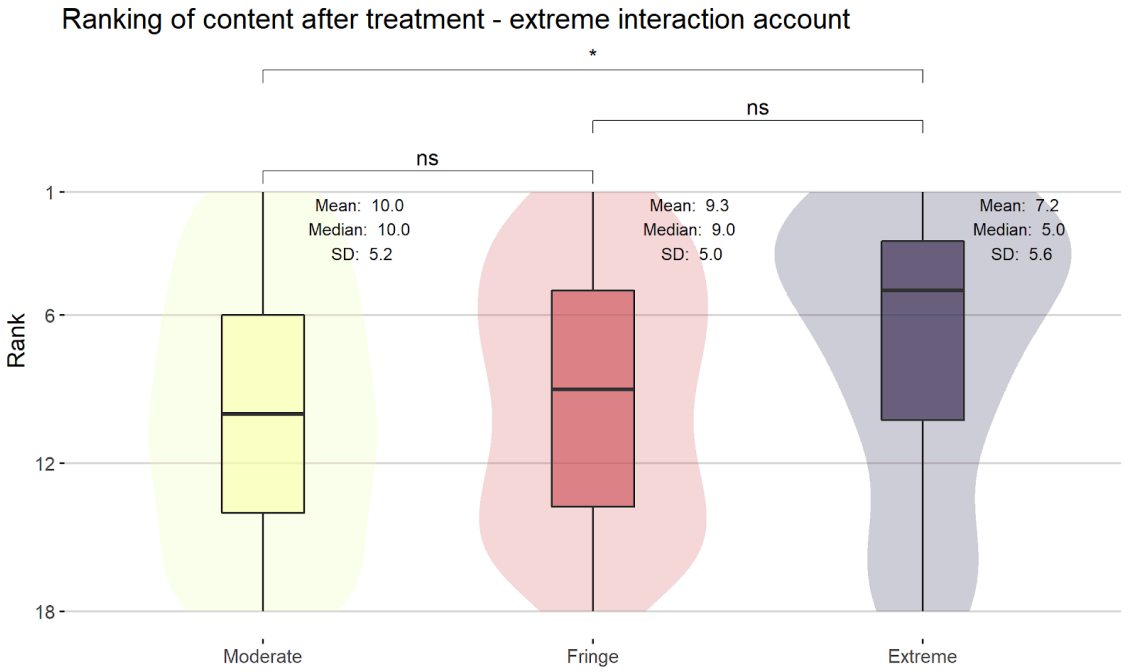

FIGURE 2: Ranking by EMI scores for the EIA condition (YouTube) and test comparisons are Wilcoxon Rank Sum Tests. * $p < 0.05$.

RQ3: Do Gab's timelines promote Extreme content?

The exploratory investigation on Gab did not yield any differences in the promotion of Extreme content in the nine observations (three timelines vs three topics). The content in the "Latest" and "Controversial" timelines showed no statistically significant differences with any of the EMI categories, and the "Popular" timeline shows a prioritisation for Fringe content above Moderate, but there is no statistically significant promotion of Extreme posts.

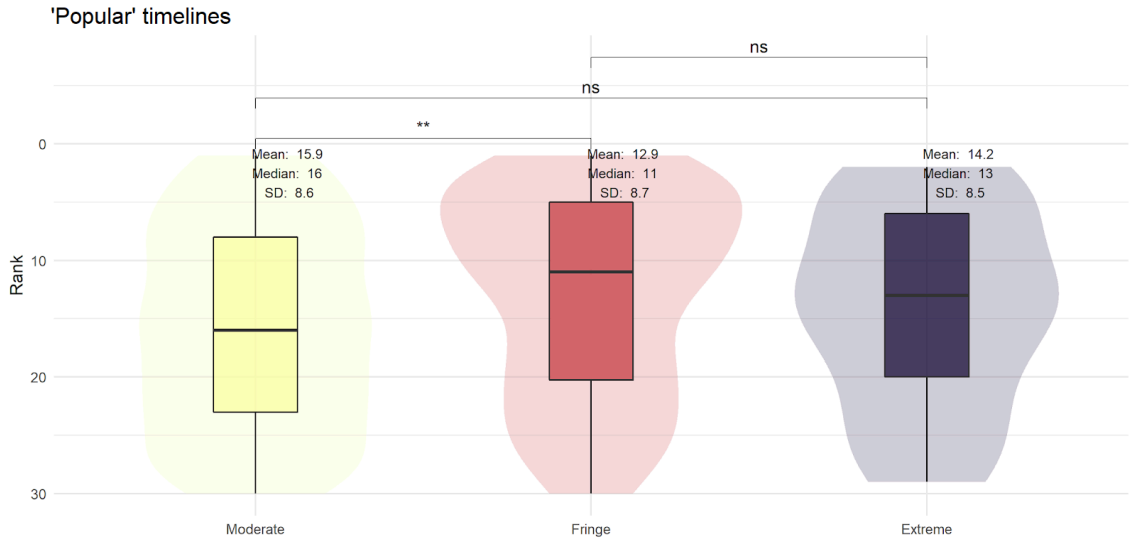

FIGURE 3: Ranking by EMI scores in the "Popular Timelines" on Gab. Test comparisons are Wilcoxon Rank sum tests. ** $p < 0.01$, * $p < 0.05$.

Policy discussion

In this section, we synthesise our empirical findings with the existing literature and policy concerns. We assess the current regulatory approaches which are addressing the problem, finding that where there is explicit legislation, it is mostly focused on algorithmic transparency and there is currently a gap in understanding on how to deal with the problem of borderline content. We argue that a move towards co-regulation between states, social media platforms, and other stakeholders may help to address some problems and concerns. This includes a lack of accountability, a homogenous approach in dealing with borderline content, and bridging knowledge gaps caused by the slow-moving pace of legislation.

Our research suggests that YouTube can amplify extreme content towards its users after a user begins to interact with far-right content. From a policy perspective, this could affirm the concerns of several stakeholders that have highlighted this problem in recent years (HM Government, 2019; Council of the European Union, 2020). RQ1 found that by applying a treatment of predominantly far-right content, users were significantly more likely to be recommended videos that were more extreme. The converse was true, too; acting predominantly with neutral content made extreme content less likely to be shown. RQ2 found that after applying an extreme treatment, far-right content was ranked higher on average than moderate. This is in keeping with the empirical literature on YouTube, which suggests that recommended videos may promote extreme content (O'Callaghan *et al.*, 2015; Schmitt *et al.*, 2018; Ribeiro *et al.*, 2019; Baugut and Neumann, 2020). Importantly,

for RQ1 and RQ2, we add a novel contribution to the wider literature by creating experimental conditions with a baseline and controls which account for personalisation, rather than relying on content that *could* be recommended to users.

It is also important to note that we find there to be minimal filtering of extreme content on both Reddit and Gab. However, we do find evidence of both extreme and fringe content on all three platforms—supported by research which posits the far-right as existing on the sites (O'Callaghan *et al*., 2015; Berger, 2018b; Lewis, 2018; Conway, Scrivens, and Macnair, 2019; Nouri, Lorenzo-Dus, and Watkin, 2019; Copland, 2020; Gaudette *et al*., 2020). For Reddit and Gab, the lack of algorithmic promotion suggests that there are other factors, possibly related to the platforms' other affordances or their user bases, that drive extreme content. To reiterate Munger and Phillips's (2019) point, it is important to consider both the "supply" and "demand" of radical content on social media platforms. This shows the interrelated nature of the debates surrounding concerns of algorithmic promotion of extremism and content removal. This, as we will expand on below, has its own set of challenges with which policymakers must deal when considering appropriate regulatory responses.

Regulatory approaches

Despite repeated concerns from policymakers, the amplification of extremist content by algorithms is currently a blind spot for social media regulation and addressing it presents challenges to legislators. One challenge is that current and planned national regulation is focused on the moderation and removal of illegal content rather than amplification. The UK Online Harms White Paper (2019) initially addressed the question of content amplification, subsequent consultations have seemingly relegated its importance. The German NetzDG (2017) avoids the issue of the amplification of extremist content through focusing on the removal of illegal hate speech content alone. Both pieces of legislation are focused on regulating the removal of harmful—or illegal in the case of NetzDG—content from social media through the implementation of heavy fines if a platform does not implement mechanisms of notice and take downs. This is mirrored in the proposed legislation from the European Parliament on preventing terrorist content online (European Parliament, 2019).

At time of writing the UK Online Harms Bill has not been presented to Parliament, therefore this information is derived from the Online Harms White Paper (HM Government, 2019) and accompanying responses. The Bill proposes a duty of care for platforms which covers terrorism and extremism. In the white paper, the UK gov-

ernment referred to the issues surrounding the operation of algorithms and their role in amplifying extreme content; "Companies will be required to ensure that algorithms selecting content do not skew towards extreme and unreliable material in the pursuit of sustained user engagement" (HM Government, 2019, p. 72). The white paper affirmed that the proposed regulator would have the power to inspect algorithms *in situ* to understand whether this leads to bias or harm. This power is analogous to the form of algorithmic auditing advocated by Mittelstadt, who argues for the "prediction of results from new inputs and explanation of the rationale behind decision, such as why a new input was assigned a particular classification" (Mittelstadt, 2016, p. 4994). This can, in principle, be implemented at each stage of the algorithm's development and lines up well with the regulator's proposed power to review the algorithm *in situ*. Impact auditing that investigates the "types, severity, and prevalence of effects of an algorithm's outputs" is also advocated which can be conducted while the algorithm is in use (Mittelstadt, 2016, p. 4995). There is also reference made to the regulator requiring companies to demonstrate how algorithms select content for children, and to provide the means for testing the operation of these algorithms, which implies the development of accountable algorithms (Kroll et al., 2016).

However, in the final government response to the consultation setting out the current plans for the Bill, algorithms are barely mentioned. In the government's final response, it is stated that search engines should ensure that "algorithms and predictive searches do not promote illegal content" referring specifically to child sexual exploitation images (HM Government, 2020, para 1.3). This has been implemented in the interim Code of Practice on child sexual exploitation images and abuse. No such reference is made in the duty of care for terrorist material or the corresponding interim Code of Practice on terrorist content and activity online.

Transparency

Where legislation does address content amplification explicitly, regulation is largely limited to transparency requirements. Recommender systems are explicitly addressed in Article 29(1) of the EU Digital Services Act (2021). This requires very large platforms which use recommender systems to set out:

> In their terms and conditions, in a clear, accessible and easily comprehensible manner, the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters that they may have made available, including at least one option

> which is not based on profiling. (2021, Art. 29(1))

The following article specifies further: where several options are available pursuant to Article 29(1) very large online platforms "shall provide an easily accessible functionality on their online interface allowing the recipient of the service to select and to modify at any time their preferred option for each of the recommender systems that determines the relative order of information presented to them". In its current iteration, Facebook's 'why am I seeing this?' tool likely meets these requirements (Sethuraman 2019). These sections only apply to very large platforms, defined as having an audience of at least 10% of the EU population (DSA 2020, para 54), and therefore smaller platforms may escape transparency requirements. Additionally, the wording of Article 29 DSA provides a large amount of discretion to social media platforms as to what parameters they choose to make available for users to modify or influence, effectively affirming the social media platforms as the leaders of recommender system regulation (Helberger, 2021). This evokes the reminder that transparency afforded to users by the 'why am I seeing this?' tool is not total transparency. The data available to users, as well as researchers, is subject to the politics of visibility and politics of knowledge within Facebook implemented via changes at an interface and software level (Stefanija & Pierson, 2020, p. 112).

Like the DSA, where the Online Harms Bill does address the regulation of algorithms, it is in pursuit of transparency. The UK government established a multi-stakeholder Transparency Working Group which included representatives from civil society and industry including Facebook, Google, Microsoft and Twitter. Discussions and recommendations in this group did not go much beyond the specifics of transparency reports (HM Government, 2020, para 2.2). One exception to this is that the recommendation reporting includes, where appropriate, information on the use of algorithms and automated processes in content moderation. However, it was noted that certain information about the operation of companies' algorithms is commercially sensitive or could pose issues if user safety is publicly disclosed (HM Government, 2020, para 6.21). The EU Counter-Terrorism Coordinator refines this point by stating that companies' transparency reports should include detailed information concerning their practices on recommendation; whether blocked or removed illegal and borderline content was promoted by the platform's algorithms. This should include several views, as well as data on how often the content was recommended to users, and whether human oversight was involved (Council of the European Union, 2020).

The German implementation of the Audiovisual Media Services Directive, the *Medienstaatsvertrag* (2020) does address the issue of content amplification but only requires media platforms such as Netflix or Amazon Prime to be heavily regulated. So-called media intermediaries, such as YouTube or other video hosting sites, are given a lighter touch. In terms of transparency, for example, media platforms must disclose the way selection criteria are weighted, the functioning of the algorithm, how users can adjust and personalise the sorting and explain the reasoning behind content recommendations (2020, pp. 78-90). By contrast, media intermediaries must disclose the selection criteria that determine the sorting and presentation of content. These disclosures must be made in easily recognisable, directly accessible, and constantly available formats (ibid, pp. 91-96). Facebook's 'Why am I seeing this?' goes further than its media intermediary obligation by disclosing the functioning of the algorithm and explaining the reasoning behind content recommendations. One interesting provision prohibits media intermediaries from discriminating against journalistic and editorial content or treating them differently without appropriate justification. If a provider of such content believes that they have been discriminated against, they can file a claim with the relevant state broadcasting authority (ibid, p. 94). It is difficult to imagine how discrimination could be proven in practice however, and it is unclear how this interacts with the objective of recommender systems, and search engines, which is to discriminate between content (Nelson and Jaursch, 2020).

While these attempts at increased transparency are welcome, it does not solve the issue of the algorithmic promotion of borderline, but legal extremist material. While a user who is presented with clear information on why a particular video was recommended to them may reconsider watching said video, they may also disregard it. This is particularly likely to happen if the recommended video is something that they are likely to find agreeable. Thus, regulation of this issue must move further than issues of transparency and into the issue of recommending borderline content. This moves into a more ethically challenging area for legislators in deciding what is considered borderline.

The case of borderline content

Policymakers have suggested that algorithms promote legal, yet borderline content that can be harmful and lead to radicalisation. As a solution, many have argued that platforms should restrict the flow of legal, yet potentially harmful content to their audiences. The EU Counter-Terrorism Coordinator notes that systems are specifically designed to target users and not just organise content generally, therefore there should be no exemption from liability (Council of the European

Union, 2020). While there is an argument that platforms such as YouTube are not the neutral intermediaries that they claim to be (Suzor, 2018), this would seem to take this argument to an impractical conclusion.

Although studies, including ours, do find there to be extreme and fringe content that within platforms' recommendations, they use neither a legal definition of extremism, nor one that mirrors terms of service. Extremism is a difficult phenomenon to define and identify and is subject to a great degree of academic debate (for example, see: Schmid, 2013; Berger, 2018a). This invariably leads to a sizable grey zone of borderline content that represents a challenge for regulation (Bishop *et al*., 2019; Vegt *et al*., 2019; Conway, 2020). In the face of such liability, over-removal is a potential problem, leading to concerns around free speech. The Coordinator attempts to address this problem by citing the platforms' ability to automatically find, limit and remove copyrighted content. However, this is clearly not a reasonable comparison given the much greater sized grey area around extreme content and the freedom of expression issues that arise.

It is unclear whether this action would be contrary to EU law. Article 14 of the e-Commerce Directive exempts intermediaries from liability for the content they manage if they fulfil certain conditions including removing illegal content as fast as possible once they are aware of its illegal nature, and that they play a neutral, merely technical and passive role towards the hosted content (e-Commerce Directive, 2000, Art. 14). The European Court of Justice takes a case-by-case approach to whether a hosting intermediary has taken a passive role, however, the use of algorithms or automatic means to select, organise or present the information would not be sufficient to automatically meet the active role standard (*Google France and Google* 2010, paras 115-120). This was echoed by the European Commission in stating that the mere fact that an intermediary hosting service provider 'takes certain measures relating to the provision of its services in a general manner does not necessarily mean that it plays an active role in respect of the individual content items it stores' (European Commission 2017, p. 11).

This approach has been criticised by *Tech Against Terrorism*, who argue that discussions of removing legal content from recommendations are misplaced and do not understand the nature of terrorists' use of the internet. They argue that it has harmful implications for the freedom of speech and the rule of law and raises serious concerns over extra-legal norm-setting. Moreover, the definitional subjectivity of concepts like extremism (as laid out above) or "harmful" result in this being difficult to operationalise (Tech Against Terrorism, 2021). They assert that norm-setting should be created by consensus-driven mechanisms which are driven by de-

mocratically accountable institutions.

Some platforms have taken steps to remove borderline content from their recommendations unilaterally. In 2017, YouTube announced that content that did not clearly violate its policies but was deemed potentially extreme (they give examples of inflammatory religious or supremacist content) would appear behind a warning and not be available for monetisation, recommendation, and not eligible for comments or endorsements (Walker, 2017). They argue that this would make the content harder to find, striking the right balance between free speech and access to information without amplifying extreme viewpoints. Facebook has adopted this approach too; they offer a range of factors that may cause them to remove content that is permitted from recommendations. Relevant for this discussion, they include accounts that have recently violated the platform's community standards, as well as accounts that are associated with offline movements that are tied to violence (Facebook, n.d.). Reddit has a policy of "quarantining" subreddits that are grossly offensive, which removes it from the platform's recommendation system and forces users to opt-in to see content (Reddit, 2021). One example of this is r/The_Donald, which Gaudette et al. (2020) identify as hosting problematic extremist content. However, the platform has been accused of only applying these measures after the subreddit had received negative media attention (Romano, 2017).

Currently, choosing to remove legal, yet potentially problematic content from recommendations is a choice for individual platforms—i.e. self-regulation. However, this can be problematic, the EU Counter-Terrorism Coordinator argues, if platforms being unable or unwilling to de-amplify content: 'Some companies have no incentive to promote a variety of viewpoints or content…since recommending polarising content remains the most efficient way to expand watch time and gather more data on customers, to better target advertising and increase the returns' (Council of the European Union, 2020, p. 5). This speaks to the lack of a strong natural coincidence between the public and private interest in this area. Absent such a natural coincidence, one or more external pressures sufficient to create such a coincidence are needed (Gunningham and Rees, 1997, p. 390). As discussed above, there are also issues with regulation which holds platforms liable for this type of content; therefore, we advocate the use of co-regulation to achieve this coincidence.

Towards co-regulation

We define co-regulation as a regulatory scheme which combines elements of self-regulation (and self-monitoring) with elements of traditional public authority and private sector elements. A key aspect of a co-regulatory regime is the self-con-

tained development of binding rules by the co-regulatory organisation and the responsibility of this organisation for these rules (Palzer, 2003). While this concept encompasses a range of regulatory phenomena, each co-regulatory regime consists of a complex interaction of general legislation and a self-regulatory body (Marsden, 2010, p 222). In the EU context it is defined as a mechanism whereby a community legislative act entrusts the attainment of the objectives defined by the legislative authority to parties which are recognised in the field.

There are opportunities to implement co-regulation in present and upcoming legislation. In the UK Online Harms Bill, the government will set objectives for the regulator's (Ofcom) code of practice in secondary legislation to provide clarity for the framework. Ofcom will have a duty to consult interested parties on the development of the codes of practice (HM Government 2020, para. 31). The requirements for consultation are low, meaning that not only social media companies but other stakeholders such as civil society can participate. However, as highlighted above, the UK Government's final consultation which sets out the current plans for the Online Harms Bill makes little mention of plans to regulate recommender systems or other content amplification; it remains unclear whether they will be addressed in the development of the codes of practice.

The DSA also provides a mechanism for very large platforms to cooperate in the drawing-up of codes of conduct, thus providing another avenue for co-regulation of this issue (DSA 2020, Art. 35). A co-regulatory scheme wherein government, industry and civil society can fully participate would be the appropriate venue to find the solution to the issue of the amplification of borderline and extremist content. This lines up with previous work on the increasingly shared responsibilities between states and companies, and the trend that social media platforms are adopting measures which are increasingly similar to administrative law (Heldt, 2019). Of particular note is the framework of cooperative responsibility sketched out by Helberger, Pierson, and Poell (2017) which emphasises the need for dynamic interaction between platforms, users and public institutions in realising core public values in these online sectors. This still leaves the issue of the ethical and practical implications of who decides what is inappropriate to be amplified but doing so in a consensual manner has clear benefits.

The primary benefit of a co-regulatory approach is the avoidance of self-regulatory or public authority regulatory approaches (Palzer, 2003). One issue with self-regulation is that there may be an accountability gap as the social media companies in question are responsible for holding themselves accountable (Campbell, 1998). Presently, that may lead to little being done to tackle the issue of amplification of

borderline and extremist content. Sufficiently bad press may provoke the company to act, but this could be subject to short-termism as the company acts in their immediate economic self-interest which could lead to hasty and arbitrary decisions (Gunningham and Rees, 1997). Thus far, initiatives to regulate social media platforms have been mostly self-regulatory such as the Global Internet Forum to Counter Terrorism, the Facebook Oversight Board, or the above-mentioned methods to remove content from recommendations.

Public authority regulatory approaches may lead to a knowledge gap, as states' attempts to regulate technology may be outdated by the time they are implemented (Ayres and Braithwaite, 1992). For example, the Online Harms Bill was proposed in 2017 and as of the writing of this paper has still not been introduced to Parliament. A co-regulatory approach provides the opportunity to implement novel solutions such as algorithmic auditing and accountable algorithms. A regulatory body such as Ofcom could work with social media platforms to conduct developmental auditing to develop adequate and sufficient safeguards in their algorithms. It could analyse the impact the algorithm has on the average user, potentially through similar methodologies as this paper, although on a much-expanded scale. Another benefit of a co-regulatory approach to this issue is that it avoids giving the responsibility for filling in the details of the law to programme developers. This avoids the problem of a programme developer designing a wide-ranging algorithm to solve a political problem, which the developer likely has little substantive expertise on, and with slight possibility of political accountability (Kroll et al., 2016). Wachter and Mittelstadt have advocated a right to 'reasonable inferences' to close the accountability gap posed by big data inferences which damage privacy, reputation, or are used in important decisions despite having low verifiability (Wachter and Mittelstadt, 2019). Should such a right be established, a co-regulatory body could audit or help design algorithms which keep inferences and subsequent recommendations, nudges and manipulations to a reasonable level.

Conclusion

We anticipate that the role of social media recommendation algorithms and extremist content will continue to be a point of policy concern moving forward. It seems inevitable that news organisations will continue to publish stories that highlight instances of unsavoury content being recommended to users and policymakers will continue to be concerned that this is harmful to users and may exacerbate radicalisation trajectories. This article has sought to provide clarity towards this future debate in two ways. Firstly, it has provided the first empirical assess-

ment of interactions between extremist content and platforms' recommendation systems in an experimental condition while accounting for personalisation. The findings suggest that one platform—YouTube—may promote far-right materials after a user interacts with it. The other two platforms—Reddit and Gab—showed no signs of amplifying extreme content via their recommendations.

Secondly, we contextualise these findings into the policy debate. At first glance, our research seems to support policy concerns regarding radical filter bubbles. However, we argue that our findings also point towards other problematic aspects of contemporary social media. More focus needs to be paid to the online radical milieu and the audience of extremist messaging. Despite repeatedly being signalled out by policymakers as problematic, there are currently few instruments in place for social media regulation. Moreover, where regulatory policy does exist, it tends to focus on transparency, which while welcome, is only one potential solution to the amplification of extreme content alone. We argue that policy is yet to fully understand the difficulties with "grey area" content as it relates to content amplification. Currently, platforms are left to self-regulate in this area and policymakers argue that they can do more. However, self-regulation can be problematic because of a lack of coincidence between public and private interests. We argue that a movement towards co-regulation offers numerous benefits as it can shorten the accountability gap while maintaining the opportunity for novel solutions from industry leaders.

Acknowledgments

We are grateful to our reviewers Amélie Heldt, Jo Pierson, Francesca Musiani, and Frédéric Dubois, who each helped improve this article through the peer-review process.

References

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons.

Ayres, I., & Braithwaite, J. (1992). *Responsive regulation: Transcending the deregulation debate*. Oxford University Press.

Azeez, W. (2019, May 15). YouTube: We're Learnt Lessons From Christchurch Massacre Video. *Yahoo Finance UK*. https://uk.finance.yahoo.com/news/you-tube-weve-learnt-lessons-from-christchurch-massacre-video-163653027.html

Bakshy, E., Messing, S., & Adamic, L. (2015). Exposure to Ideologically Diverse News and Opinion on Facebook. *Science Express*, 1–5. https://doi.org/10.1111/j.1460-2466.2008.00410.x

Baugut, P., & Neumann, K. (2020). Online propaganda use during Islamist radicalization. *Information Communication and Society*, *23*(11), 1570–1592. https://doi.org/10.1080/1369118x.2019.1594333

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

Berger, J. M. (2018a). *Extremism*. MIT Press.

Berger, J. M. (2018b). *The Alt-Right Twitter Census*.

Bishop, P. (2019). *Response to the Online Harms White Paper*. Swansea University, Cyber Threats Reseacxh Centre. https://www.swansea.ac.uk/media/Response-to-the-Online-Harms-White-Paper.pdf

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227. https://doi.org/10.1007/s10676-013-9321-6

Bruns, A. (2019). Filter bubble. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1426

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Campbell, A. J. (1998). Self-regulation and the media. *Fed. Comm. LJ*, *51*.

Christchurch Call. (2019). *The Call*. https://www.christchurchcall.com/call.html

Commission for Countering Extremism. (2019). *Challenging Hateful Extremism*.

Conway, M. (2016). Violent Extremism and Terrorism Online in 2016. *The Year in Review', Vox Pol*.

Conway, M. (2020). Routing the Extreme Right: Challenges for Social Media Platforms'. *RUSI Journal*. https://doi.org/10.1080/03071847.2020.1727157

Conway, M., Scrivens, R., & Macnair, L. (2019). *Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends'*. ICCT Policy Brief.

Copland, S. (2020). Reddit quarantined: Can changing platform affordances reduce hateful material online?'. *Internet Policy Review*, *9*(4), 1–26. https://doi.org/10.14763/2020.4.1516

Council of the European Union. (2020). *The Role of Algorithmic Amplification in Promoting Violent and Extremist Content and its Dissemination on Platforms and Social Media*.

Courtois, C., Slechten, L., & Coenen, L. (2018). Challenging Google Search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. *Telematics and Informatics*, *35*(7), 2006–2015. https://doi.org/10.1016/j.tele.2018.07.004

DeVito, M. A. (2016). From Editors to Algorithms. *Digital Journalism*, 1–21. https://doi.org/10.1080/21670811.2016.1178592

Dylko, I. (2018). Impact of Customizability Technology on Political Polarization. *Journal of Information Technology and Politics*, *15*(1), 19–33. https://doi.org/10.1080/19331681.2017.1354243

Eslami, M. (2015). "I always assumed that I wasn't really that close to [her]". *CHI'15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–162. https://doi.org/10.1145/2702123.2702556

E.U. Commission. (2000). *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services*. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031.

E.U. Commission. (2017). *Tackling Illegal Content Online; Towards an enhanced responsibility of online platforms*. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=47383

E.U. Commission. (2020). *Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC*. https://www.euractiv.com/wp-content/uploads/sites/2/2020/12/Digital_Services_Act__1__watermark-3.pdf.

European Parliament. (2019). *Terrorist content online should be removed within one hour, says EP* [Press Release]. European Parliament. https://www.europarl.europa.eu/news/en/press-room/20190410IPR37571/terrorist-content-online-should-be-removed-within-one-hour-says-ep.

Facebook Help Centre. (nd). *What are recommendations on Facebook?* https://www.facebook.com/help/1257205004624246

Gaudette, T. (2020). Upvoting Extremism: Collective identity formation and the extreme right on Reddit. *New Media and Society*. https://doi.org/10.1177/1461444820958123

Government, H. M. (2020). *The Government Report on Transparency Reporting in relation to Online Harms* [Report]. The Stationary Office. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/944320/The_Government_Report_on_Transparency_Reporting_in_relation_to_Online_Harms.pdf

Gunningham, N., & Rees, J. (1997). Industry self-regulation: An institutional perspective. *Law & Policy*, *19*(4), 363–414. https://doi.org/10.1111/1467-9930.t01-1-00033

Haim, M., Graefe, A., & Brosius, H. B. (2018). Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News'. *Digital Journalism*, *6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, *34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Heldt, A. (2019). Let's Meet Halfway: Sharing New Responsibilities in a Digital Age'. *Journal of Information Policy*, *9*, 336–369. https://doi.org/10.5325/jinfopoli.9.2019.0336

H.M. Government. (2019). *Online Harms White Paper* [White Paper]. The Stationary Office. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf.

Holbrook, D. (2015). Designing and Applying an "Extremist Media Index". *Perspectives on Terrorism*, *9*(5), 57–68. https://doi.org/10.1088/0031-9155/49/9/004

Holbrook, D. (2017a). Terrorism as process narratives: A study of pre-arrest media usage and the emergence of pathways to engagement. In *Terrorism and Political Violence, In Press*.

Holbrook, D. (2017b). *What Types of Media Do Terrorists Collect?* International Centre for Counter-Terrorism.

Kraska-Miller, M. (2013). *Nonparametric Statistics for Social and Behavioral Sciences*. CRC Press.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable

algorithms. *U. Pa. L. Rev*, *165*, 633.

Ledwich, M., & Zaitsev, A. (2019). *Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization*. http://arxiv.org/abs/1912.11211.

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

Marsden, C. T. (2010). *Net Neutrality: Towards a Co-regulatory Solution*. Bloomsbury Academic.

Mittelstadt, B. (2016). Automation, algorithms, and politics| auditing for transparency in content personalization systems. *International Journal of Communication*, *10*, 12.

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

Munger, K., & Phillips, J. (2019). *A Supply and Demand Framework for YouTube Politics Introduction to Political Media on YouTube*. Penn State Political Science.

Munger, K., & Phillips, J. (2020). Right-Wing YouTube: A Supply and Demand Perspective. *International Journal of Press/Politics*. https://doi.org/10.1177/1940161220964767

Napoli, P. M. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication Theory*, *24*(3), 340–360. https://doi.org/10.1111/comt.12039

Napoli, P. M. (2015). Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. *Telecommunications Policy*, *39*(9), 751–760. https://doi.org/10.1016/j.telpol.2014.12.003

Nelson, M., & Jaursch, J. (2020). Germany's new media treaty demands that platforms explain algorithms and stop discriminating. Can it deliver? *Algorithm Watch*. https://algorithmwatch.org/en/new-media-treaty-germany/

Nouri, L., Lorenzo-Dus, N., & Watkin, A. (2019). Following the Whack-a-Mole Britain First's Visual Strategy from Facebook to Gab. *Global Research Network on Terrorism and Technology*, *4*.

O'Callaghan, D. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Ottoni, R. (2018). Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination'. *Proceedings Ofthe 10th ACM Conference on Web Science*. https://doi.org/10.1145/3201064.3201081

Palzer, C. (2003). Self-monitoring v. Self-regulation v. Co-regulation. In W. Closs, S. Nikoltchev, & European Audiovisual Observatory (Eds.), *Co-regulation of the media in Europe* (pp. 29–31).

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Reddit Help. (2021). *Quarantined Subreddits*. https://reddit.zendesk.com/hc/en-us/articles/360043069012-Quarantined-Subreddits

Ribeiro, M. H. (2019). *Auditing Radicalization Pathways on YouTube*. ACM Symposium on Neural Gaze Detection. http://arxiv.org/abs/1908.08313.

Ricci, F., Rokach, L., & Shapira, B. (2011). *Recommender Systems Handbook*. Springer. https://doi.org/10.1007/978-0-387-85820-3

Romano, A. (2017). Reddit's TheRedPill, notorious for its misogyny, was founded by a New Hampshire state legislator. *Vox*. https://www.vox.com/culture/2017/4/28/15434770/red-pill-founded-by-robert-fisher-new-hampshire

Schmid, A. P. (2013). *Radicalisation, De-Radicalisation, Counter-Radicalisation: A Conceptual Discussion and Literature Review* [Research Paper]. International Centre for Counter-Terrorism. http://www.icct.nl/download/file/ICCT-Schmid-Radicalisation-De-Radicalisation-Counter-Radicalisation-March-2013.pdf.

Schmitt, J. B. (2018). Counter-messages as prevention or promotion of extremism?! The potential role of YouTube. *Journal of Communication*, *68*(4), 758–779. https://doi.org/10.1093/joc/jqy029

Seaver, N. (2018). Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*. https://doi.org/10.1177/1359183518820366

Sethuraman, R. (2019). *Why Am I Seeing This? We Have an Answer for You*. Facebook. https://about.fb.com/news/2019/03/why-am-i-seeing-this/

Stefanija, A. P., & Pierson, J. (2020). Practical AI Transparency: Revealing Datafication and Algorithmic Identities. *Journal of Digital Social Research*, *2*(3), 84–125.

Sunstein, C. R. (2002). The law of group polarization. *The Journal of Political Philosophy*, *10*(2), 175–195. https://doi.org/10.1002/9780470690734.ch4

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society*, *4*(3). https://doi.org/10.1177/2056305118787812

Tech Against Terrorism. (2021). *Content personalisation and the online dissemination of terrorist and violent extremist content* [Position paper]. https://www.techagainstterrorism.org/wp-content/uploads/2021/02/TAT-Position-Paper-content-personalisation-and-online-dissemination-of-terrorist-content1.pdf

van Der Vegt, I. (2020). Online influence, offline violence: Language Use on YouTube surrounding the "Unite the Right" rally. *Journal of Computational Social Science*, *4*, 333–354. https://doi.org/10.1007/s42001-020-00080-x

van Der Vegt, I., Gill, P., Macdonald, S., & Kleinberg, B. (2019). *Shedding Light on Terrorist and Extremist Content Removal* (Paper No. 3). Global Research Network on Terrorism and Technology. https://rusi.org/explore-our-research/publications/special-resources/shedding-light-on-terrorist-and-extremist-content-removal

Vīķe-Freiberga, V., Däubler-Gmelin, H., Hammersley, B., & Maduro, L. M. P. P. (2013). *A Free and Pluralistic Media to Sustain European Democracy* [Report]. EU High Level Group on Media Freedom and Pluralism.

Wachter, S., & Mittelstadt, B. (2019). A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review*, *2*, 494–620. https://doi.org/10.7916/cblr.v2019i2.3424

Walker, K. (2017, June). Four Steps We're Taking Today to Fight Terrorism Online [Blog post]. *Google*. https://www.blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/

Waters, G., & Postings, R. (2018). *Spiders of the Caliphate: Mapping the Islamic State's Global Support Network on Facebook* [Report]. Counter Extremism Project. https://www.counterextremism.com/sites/default/files/Spiders%20of%20the%20Caliphate%20%28May%202018%29.pdf

Whittaker, J. (2020). Online Echo Chambers and Violent Extremism. In S. M. Khasru, R. Noor, & Y. Li (Eds.), *The Digital Age, Cyber Space, and Social Media: The Challenges of Security & Radicalization* (pp. 129–150). Dhaka Institute for Policy, Advocacy, and Governance.

Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, *5*(1). https://doi.org/10.14763/2016.1.401

Statutes cited:

Bundestag (2017), Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG).

Bundestag (2020), State Treaty on the modernisation of media legislation in Germany (Medienstaatsvertrag) https://ec.europa.eu/growth/tools-databases/tris/en/index.cfm/search/?trisaction=search.detail&year=2020&num=26&dLang=EN

Filter Bubble Transparency Act. (2019). S, LYN19613

French National Assembly. (2019). Lutte Contre la Haine sur Internet.

Cases cited:

Case C-236/08 Google France SARL and Google Inc. v Louis Vuitton Malletier SA [2010] R.P.C. 19

Published by
Alexander von Humboldt Institute for Internet and Society

in cooperation with
CREATe
cnrs centre internet et societe
R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya
University of Tartu Johan Skytte Institute of Political Studies