Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 5.1 (2023) 109–140
HTTPS://DOI.ORG/10.5117/CCR2023.1.005.ZIER

Algorithmic Recommendations' Role for the Interrelatedness of Counter-Messages and Polluted Content on YouTube – A Network Analysis

Lisa Zieringer
Department of Media and Communication, LMU Munich

Diana Rieger
Department of Media and Communication, LMU Munich

Abstract
Counter-messages are used by civil education, youth prevention actors, and security agencies to counter the magnitude of polluted content. On the Internet, algorithmic operations of intermediaries affect how users encounter and receive polluted content. As counter-messages often show similar keywords, algorithms establish connections between counter-messages and polluted content, primarily because they share mutual topics. Against the background of legislative attempts to stop the spread of extremist online content, this paper aims to further investigate the interrelatedness of counter-messages and polluted content on YouTube due to the platform's recommendation algorithm. To that end, two information network analyses were conducted based on each five seed videos of two differently designed counter-message campaigns one year after their publication on YouTube in 2019. Five thousand four hundred of the 35,982 videos of the two networks were analyzed qualitatively and manually. Results show that counter-messages are indirectly strongly connected to more polluted content. We further identify the campaigns' design and setup on YouTube as factors that can cause the interrelatedness between counter-messages and polluted content.

Keywords: Information network analysis; YouTube; Algorithms; Counter-messages; Polluted content

Introduction

Extremist actors who use social media to rapidly distribute their ideas to a large audience and easily reach young audiences (Schmitt et al., 2018) exacerbate worldwide societal concerns on the Internet as a multiplier of extremist thoughts and ideas (Gottfried & Shearer, 2016). Most adolescents, and almost two-thirds in most nations, had been exposed to online hate in

the previous three months (Reichelmann et al., 2021). In Germany, more than three-quarters (77%) of Internet users have experienced hate speech (Landesanstalt für Medien NRW, 2022), and half of them have encountered conspiracy narratives (e.g., denial of human-made climate change; Sängerlaub & Schulz, 2021), and 37% of 14 to 19-year-olds have come across extremist content at least sometimes in their online environment (Nienierza et al., 2021). Although hate speech or conspiracy narratives do not necessarily contain extremist ideas, they can be regarded as indicators of radicalization dynamics in the online environment (Schulze et al., 2022): Therefore, we coin extremist content, hate speech, and conspiracy narratives as 'polluted content' following Wardle's theoretical conceptualization of a "polluted information ecosystem" (2018, p. 951). When analyzing how Internet users find polluted content, several routes are feasible. Aside from actively seeking it, most users report more passive encounters, such as being forwarded videos and, most often, simply 'stumbling' upon it (Costello et al., 2016; Reinemann et al., 2019; Rieger et al., 2013).

In this paper, we depart from the notion that ideological biases in algorithmic recommendations (Haroon et al., 2022) may influence the probability of encountering and processing polluted content. First, algorithmic recommendations via social media may encourage people to concern themselves with content from attitude-inconsistent sources they would usually not engage with (Messing & Westwood, 2014). Second, younger users could be exposed to content they cannot process properly, as they might lack a critical reflection of content and sources (Morris, 2016; Sonck et al., 2011). Moreover, "in the context of extremist content, algorithmic 'recommendation' could disguise ideological partisanship as they make content appear 'related'" (Schmitt et al., 2018, p. 781). This might be especially important to consider for another reason: Prevention actors put counter-messages in the environment of polluted content to counter the promotion of extremist ideologies and ideas (Schmitt et al., 2018): Counter-messages "can be defined as positive messages directed against extremist ideologies, core elements of ideologies, or violent extremist behavior" (Schmitt et al., 2018, p. 783).

Counter-message campaigns are often published on popular social media channels like YouTube. These often share similar keywords with their counterpart of videos containing polluted content, directly referring to extremist actors by their names or organizations (Schmitt et al., 2018). Prevention actors are thus confronted with a potential structural problem due to the interconnectedness of counter-messages (CM) and polluted content (PC): Watching one CM video could direct users to videos containing PC.

Exposure to PC on YouTube may lead to a self-reinforcing process of diving into extremist parts of the Internet (O'Callaghan et al., 2015). The case of YouTube might be critical as it represents one of the most frequently used social media platforms (Newman et al., 2022): Adolescents, in particular, are heavy users of YouTube and the primary target group of propaganda (Gottfried & Shearer, 2016; Schmitt et al., 2018). Further, internal referrals on YouTube are the main factor leading to exposure to videos besides active search (Zhou et al., 2010). Thus, YouTube's recommendation system is a crucial mechanism through which users engage with content (Figueiredo et al., 2011).

Schmitt et al. (2018) analyzed the networks of two German CM campaigns on YouTube. They concluded that counter-messages are closely or directly related to "problematic, extremist" content. Their data collection occurred before the so-called Network Enforcement Act, aiming at regulating and deleting PC, was introduced in Germany in June 2017. The current study, therefore, aims at achieving two objectives: 1) It is designed to provide a conceptual replication of Schmitt et al.'s (2018) network study to investigate whether the findings replicate to other newer German anti-extremism campaigns. 2) The second goal refers to recent attempts to work against PC online. On the European level, the Code of Conduct tries to coordinate with big platform operators. The Digital Services Act aims to mitigate systemic risks, such as manipulation and disinformation, by introducing accountability frameworks and transparency rules. In Germany, a (debated) law, the Network Enforcement Act, is meant to fight unlawful content on the Internet, including hate speech and misinformation, by obligating social networks with more than two million registered users in Germany to remove manifestly / other unlawful content from their platforms within 24 hours / seven days of a complaint. Therefore, we also tested the closeness of relations between CM and PC due to YouTube's recommendation algorithm three years after the law's introduction.

Theoretical Background

Polluted content: Extremist messages, conspiracist videos, and hate speech

In their network analysis on YouTube, Schmitt et al. (2018, p. 782) refer to "messages used to promote extremism [and distinguish between] "different 'problematic' shapes, which may overlap each other, such as: (a) hate speech, (b) conspiracy theories, and (c) propaganda." Wardle (2018) points out the

need for a more thorough theoretical grounding and offers a theoretical conceptualization of a polluted information ecosystem: She differentiates between seven categories of information disorder and distinguishes false and harmful messages along the three types of mis-, dis-, and malinformation. Misinformation is defined as false content not intended to harm, dis-information as false content intended to harm, and mal-information as truthful content intended to harm. However, incitements to violence, hate speech, and extremist propaganda are difficult to research solely through the lens of information disorder because these contents may be legal to legislators in other contexts, causing harm to individuals, organizations, or democracy (Wardle, 2018).

Accordingly, we seek to heed the call for shared definitions and build on this framework (Wardle, 2018) as well as relate this study to previous work (Schmitt et al., 2018): We base our study on counter-message campaigns, which aim at working against polluted content, the collective term we use to refer to forms of hate speech and conspiracy narratives as well as extremist messages. Nienierza et al. define extremism as "a political worldview directed against the democratic constitutional state's fundamental values and core principles such as the equality of men, basic human rights, mutual respect, and the rule of law" (2019, p. 2). It can further be described as strategic communication (Arnold, 2003) that systematically aims to convince its audience of an ideology. It is often differentiated into left-wing, right-wing, or religiously motivated extremism in Germany (Reinemann et al., 2019).

Extremists publish audiovisual material, often termed extremist online propaganda, to reach their audience (Frischlich et al., 2018). Schmitt et al. (2018) differentiate extremist propaganda between right-wing and Islamist extremist propaganda and can verify close links to counter-messages. Hate speech can be described as "norm-transgressing communication that is [...] characterized by the derogation and defamation of [...] members of targeted social groups" (Rieger et al., 2018, p. 461). It can contain insults and abusive language, be represented by explicitly racist or sexist insults and incitement to violence but also manifest itself more implicitly (Gagliardone et al., 2015; Rieger et al., 2021). Conspiracy narratives can be described as "a proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons – the conspirators – acting in secret" (Keeley, 1999, p. 116). Over time "conspiracy theories have flourished on social media, raising concerns that such content is fueling the spread of disinformation, supporting extremist ideologies, and in some cases, leading to violence" (Faddoul et al., 2020, p. 1). Conspiracy

narratives do not always need to be communicated with the purpose to harm, and so might comprise both disinformation and misinformation (Santos-d'Amorim & Miranda, 2021). In January 2019, YouTube announced it would modify its recommendation algorithm to downgrade conspiracist videos (YouTube, 2019).

Countering polluted content with counter-messages

Civil education, prevention actors, and security agencies strive to counter polluted content to prevent its potential influence. Counter-messages constitute one part of their prevention activities. There are numerous CM formats, ranging from visual (texts and graphics) to audio (speeches and podcasts) to audio-visual (videos), with videos being the most employed by prevention actors (e.g., "ExitUSA" launched by the American non-profit organization Life After Hate). Counter-messages are known in research focusing on countering and preventing violent extremism as the prevention of new radicalization processes, particularly among adolescents (Caplan & Caplan, 2000).

Studies on the effectiveness of CM in counteracting the possible effects of polluted content are mixed (Frischlich et al., 2018; Hemmingsen & Castro, 2017). Two experimental studies demonstrate that counter-messages displayed directly before polluted content can decrease the evaluation of PC (Frischlich et al., 2018). Further, Counter-messages seem most effective when they argue openly, in a two-sided manner, which does not evoke reactance (Schmitt et al., 2021; Braddock, 2022). However, the interaction of PC and CM on the Internet must be studied in conjunction with the platforms' algorithmic operations, which impact users' perceptions of online material and selections (Saurwein et al., 2015).

Algorithmic curation of online information as a potential source of bias

To better understand exposure to PC, research on selective exposure has revealed that people with extreme attitudes are more likely than those with moderate attitudes to expose themselves to content selectively (Stroud, 2010) and also show more certainty in their beliefs (Wojcieszak, 2009). However, algorithms also play a vital role in selection processes (Schmitt et al., 2018; Thorson et al., 2019): They entail essential gatekeeping functions by making selection decisions for aggregators, search engines, or social media

(O'Callaghan et al., 2015). Often, algorithmic selections increase the visibility of latent biases within data sets and biases of programmers (Rainie & Anderson, 2017) as well as human behaviors (Saurwein et al., 2015) and nonconscious biases in human thinking. Algorithmic recommendations can influence the content users consume, the diversity of users' exposure, and potentially reinforce (some) users' biases from selective exposure. Individual, self-selected, and algorithmic personalization can lead to information cocoons (Zuiderveen Borgesius et al., 2016) or filter bubbles (Pariser, 2011).

Recent theory advancements include a distinction between supply, exposure, and consumption diversity when investigating recommender systems and (their effects on) media diversity (Loecherbach et al., 2020) to account for both individual news selection mechanisms and algorithmic selection and ordering (Mattis et al., 2022). The concept of filter bubbles has been theoretically (Dahlgren, 2021) and empirically challenged: Results of empirical studies (e.g., Flaxman et al., 2016; Krafft et al., 2019; Nguyen et al., 2014) and a synthesis of empirical research on the extent and effects of self-selected and pre-selected (i.e., algorithmic) personalization (Zuiderveen Borgesius et al., 2016) show little evidence of strong personalization effects regarding (news) recommendations. Research suggests it depends on the platform and the way it is used: Knudsen (2022) conducted two online experiments that simulate different news recommender systems and unobtrusively logged user behavior to find that the increase or decrease of the chance that selective exposure occurs depends on what the news recommender system "is designed to achieve." Jürgens and Stark (2022) use a four-month tracking dataset and a comprehensive content analysis covering the online news consumption of over 10,000 German citizens to show that short-term usage of platforms uniformly increases exposure diversity, whereas long-term reliance can lead to decreases.

Regarding YouTube, one of the most frequently used social media platforms (Newman et al., 2022), results from a sentiment and social network analysis investigating users' expressed opinions on three different political topics provide evidence for a "moderate level of connections between dissimilar YouTube comments but few connections between agreeing comments" (Röchert et al., 2020, p. 81). Haroon et al. (2022) conducted a systematic audit of YouTube's recommendation system using a hundred thousand sock puppets, combining research on YouTube's recommendation algorithm with an investigation of ideological bias, its magnitude, and radicalization (progressive extremity of recommendations). The findings reveal that YouTube recommendations steer users, particularly right-leaning users, to ideologi-

cally biased and increasingly extremist content. Relatedly, regarding promoting conspiracist videos on YouTube, Faddoul and colleagues (2020) built a classifier for automatically detecting conspiracy narratives and identified a positive correlation between the source video's conspiracy likelihood and the recommended video's conspiracy likelihood.

The role of YouTube's algorithms for the interrelatedness of CM and PC

YouTube defines related videos as those "a user is likely to watch after having watched the given seed video" (Davidson et al., 2010, p. 294). The relatedness of videos on YouTube is based on the interconnectedness of channels, producers, videos, similar catchphrases, user data and data of related or similar users (Davidson et al., 2010; Covington et al., 2016). The thematic congruence of videos, indicated by mutual keywords or tags, for example, "Islam", can lead the recommendation algorithm to link videos of a contrary message. This opens a new viewpoint for research on personalization, as such linkages could result in a broader set of attitudes and perspectives (Bode & Vraga, 2015). The interconnectedness of counter-messages and polluted content could be high due to the thematic congruence of videos. This may lead to a promotion of PC through exposure to CM. Polluted content might then as well promote counter-messages because they could be perceived as "related", but it needs to be considered that there might already be more published propaganda material than CM (Schmitt et al., 2018).

Ledwich and Zaitsev (2019) investigated the role of YouTube algorithms in suggesting radical content. Their classification of almost 800 political channels and a study of the suggestions obtained by each channel type indicate that YouTube's recommendation algorithms actively dissuade consumers from visiting extreme or extremist content. In contrast, the empirical findings of an auditing study of YouTube suggestions demonstrate strong ideological bias in recommendations based on the user's prior exposure: "The number of biased videos at higher depths is not only greater but [...] the recommended videos are also increasingly radical" (Haroon et al., 2022, p. 17).

Both studies deal with PC distribution on YouTube and come to very different conclusions. However, they do not consider the role of counter-messages, i.e., their objective to counter PC while bearing the risk of promoting PC. Considering this inconsistency, the research gap regarding the inclusion of CM, and the pioneer study of Schmitt et al. (2018), the following research questions are aimed at investigating the relatedness of CM and PC

on YouTube after the implementation of the Network Enforcement Act: RQ1: How closely are German CM related to PC on YouTube? RQ2: How closely are German CM related to other CM on YouTube?

Method

Description of seed videos

Two exemplary German CM campaigns published on YouTube in 2019 have been selected as they (a) represent successful, nationwide CM web video projects, (b) are coherent and comprehensive regarding their content, (c) address two different topics: (1) countering Islamist extremism, and (2) countering racism, both right-wing and Islamist extremism, and discrimination of people identifying as LGBT+[1], and (d) concentrate on different target groups (cp. Schmitt et al., 2018):

The campaign *Jamal al-Khatib* (JAK) (1), launched by the non-profit organization Turn – Association for the Prevention of Violence and Extremism, aims at conveying alternative narratives to jihadist propaganda. The web video project targets young people susceptible to jihadist or Salafist online propaganda and those who already sympathize with jihadist or Salafist groups and whose online lives are dominated by extremist ideas. With the campaign's second season of CM, five videos have been published on the eponymous YouTube channel *Jamal al-Khatib*. With these five videos the creators explain selected concepts, such as takfir[2], honor, shirk[3], democracy, and resistance. The overall CM campaign is characterized by its narrative approach, as a fictive male protagonist tells his stories of how he managed to exit extremist groups (Jamal al-Khatib, 2020).

The campaign *Say My Name* (SMN) (2), which is produced by "Kooperative Berlin" on behalf of the German Federal Agency of Civic Education, contains nine videos in which six female German YouTubers talk about their religion, ethnic origin, and sexual identity. Discrimination due to these human identifiers is a key theme the creators address by discussing their or friends' experiences. The campaign aims at strengthening the importance of democracy and plurality for living together in society and is targeted at young women between 14 and 25 years (Say My Name, 2020).

Data collection procedure

For the data analysis, we selected all five videos of the second season of the JAK campaign and five out of the nine SMN videos to achieve comparable results between the two campaigns. The five SMN videos with the highest

number of views have been chosen, as this selection criterion indicated societal relevance regarding media contact and potential effects. The total number of views of the SMN campaign (n = 110,430) is comparable to the number of views of the JAK campaign (n = 147,465), reached in February 2020, a year after the videos' publication in 2019. The five videos of each campaign represent the seeds for data collection.[4]

The online tool YouTube Data Tools – Video Network Module (Rieder, 2015) was used to scrape relevant network data from YouTube's application programming interface endpoint (Google Developers, 2017). These network data include a list of "related videos" and their metadata (e.g., video and channel ID & title) for each list of seeds. Starting with the five seeds for each campaign, related videos, videos "a user is likely to watch after having watched the given seed video" (Davidson et al., 2010, p. 294), were retrieved from the search/list#relatedToVideoId API endpoint. The tool captures a maximum of 50 'related videos' per item. Crawl depth was set to "2" and starts with "0", representing the relations between the seeds. Data collection occurred from the 30th of January to the 1st of February 2020. Aiming at reduced biased results due to the researchers' search history and behavior (see also Schmitt et al., 2018), browser history, download history, cookies and other website data, images and cache data, passwords, autofill data, website settings and hosted app data have been deleted.

Network analysis

The main objective of using network analysis is to extract relational data of entities instead of investigating independent objects. The analytical concept of a network can be defined as a set of entities, such as people, groups, or videos, with some interactions or relationships between them. For this study, a network was created for each of the two CM campaigns. The software Gephi (Version .9.2; Bastian et al., 2009) was used to visualize the data, whereby nodes represent the entities, namely CM videos and edges stand for the relationships between videos, namely links to related videos.

To extract the community structure of the two networks, modularity optimization was used: The heuristic method not only outperforms other community detection methods in terms of computation time, but the algorithm also generates high-quality communities, as measured by the so-called modularity, and constitutes a built-in algorithm in Gephi (Blondel et al., 2008). The networks' nodes and edges were clustered into communities based on modularity measures. The modularity of a partition refers to a value between [-1;1] that measures the density of links inside communities

compared to links between communities (Blondel et al., 2008). Nodes are likely to share similar characteristics within a community (as opposed to across communities). Modularity is used to assess the quality of clustering (Newman & Girvan, 2004), whereby the values of the modularity measure may vary between zero and one. A value $M < .4$ indicates a low distinctiveness of clusters, a value between $M = .4$ and $M < .6$ hints at a medium level of distinctiveness, and a value $M \geq .6$ represents high separation among the clusters (Himelboim et al., 2013).

ForceAtlas2, a force-directed layout algorithm, was applied to each data set to visualize the data. Simulating a physical system, the layout exhibits the spatial structure of the concerning network as "nodes repulse each other like charged particles, while edges attract their nodes, like springs" (Jacomy et al., 2014, p. 2). The final network graph's visualization facilitates data analysis and interpretation. Nodes linked by various edges are located in the same part of the network, and nodes with fewer connections to other nodes are further apart. The "very essence is to turn structural proximities into visual proximities, facilitating the analysis and in particular the analysis of social networks" (Jacomy et al., 2014, p. 2).

The Eigenvector centrality of each node has been calculated to analyze the significance of each node in the two networks. Eigenvector centrality measurement represents the centrality of a node regarding the global structure of the network and assigns relative scores to all videos in the network (Al-Taie & Kadry, 2017). Relations to nodes with high scoring contribute more to the score of the respective node than relations to nodes with low scoring. The more a node is connected to other well-connected nodes, the higher is the Eigenvector centrality value and thus the influence of this node.

Content Analysis

While the networks' analysis is based on the entire dataset, the content analysis of the communities is based on a sample: More specifically, a randomized sample of 15% of all videos was drawn for each cluster of the two networks. Based on the approach of Schmitt and colleagues to "get a sample size small enough to work with [...] but balanced and big enough to be representative for the total cluster" (2018, p. 802), it was decided for a 15% sample, which constitutes a scope of 5,397 videos in total. After data collection and sample drawing, each of the 5,400 videos (three more due to rounding errors when drawing the sample individually for each cluster) and its metadata were qualitatively and manually analyzed. The quality of the scraped meta data of the 5,400 videos, namely their titles, associated keywords, and descriptions,

strongly depends on what the uploader provides.

The labeling process of the YouTube videos was affected by different elements of the videos: Every video was watched for at least several seconds, and every video's metadata, title, description, as well as comments were scanned. In the case of uncertainty or ambiguities, videos were watched longer, and further information was sought (e.g., researching names and organizations or using translator tools in case of language barriers).

Standard YouTube categories such as "Autos & Vehicles" or "Comedy" (for a detailed listing, see Table 2) were used to characterize the videos, as recommended by prior studies (Filippova & Hall, 2011; Schmitt et al., 2018). On top of these categories, a new one, "Religion & Religious Music", was included to fill a gap in the nomenclature. Based on previous research (O'Callaghan et al., 2015; Schmitt et al., 2018), polluted content was operationalized with four different categories, namely Islamist extremist (IE) propaganda, right-wing extremist (RE) propaganda, hate speech, and conspiracist videos. Following Schmitt and colleagues (2018), propaganda was differentiated into the widely spread forms of right-wing extremist and Islamist extremist propaganda.

Right-wing extremism is understood as an "ideology that encompasses authoritarianism, anti-democracy and exclusionary and/or holistic nationalism" and thus messages entailing authoritarianism, nationalism, racism, anti-Semitism, xenophobia, antidemocracy or populism were classified as RE propaganda (Carter, 2018, p. 157). Islamist extremism can also be described as political extremism, but the relation to religion plays the most important role (Reinemann et al., 2019): A religion itself becomes an extremist trait if politics and society are to be shaped strictly according to the standards of a single true religion – in Germany and Europe, the relation between Islam and Islamism is particularly strong in public discourse. Messages referencing Islamist language, logos, symbols, organizations, or ideologies were classified as Islamist extremist messages. In contrast to extremist propaganda, hate speech does not necessarily serve a political or religious goal (see Rieger et al., 2018).

Messages were classified as hate speech when they were targeted at individuals or groups associated with criteria such as race or gender if these were not considered to be derived from a political or religious ideology. Conspiracist videos were classified as such when messages included the explanation of events "in terms of the significant causal agency of [...] conspirators" (Keeley, 1999, p. 116). Counter-messages were coded as such if the video's primary purpose was to actively counter any form of PC. It is noteworthy that almost every counter-message revealed an institutional

affiliation.

To assess the coding reliability of the individual categories, YouTube videos from all clusters of the two networks were reprocessed to compare the first and second coding two years later. As the content analysis was conducted by one coder, intracoder reliability, particularly Cohen's Kappa, was regarded as a suitable measurement (Gwet, 2008). The analysis of 300 double-coded videos reveals substantial agreement (Cohen's Kappa = .72).

Results

Networks' key metrics

The network of the JAK campaign is composed of 16,890 nodes and 359,234 directed edges. The average number of edges between the nodes in the network (average degree) amounts to 21.269. The ratio of the number of relations present to the maximum number of relations possible (network density, [0;1]) equals .001. The network of the campaign SMN contains more nodes (19,092) and more edges (488,561), the average degree is higher (25.590), and the network density is comparable (.001) to the JAK campaign.

	Jamal al-Khatib	*Say My Name*
Number of selected seeds	5	5
Number of total views of the seeds	147,465	110,430
Number of nodes (videos)	16,890	19,092
Number of (directed) edges (links)	359,234	488,561
Average degree	21.269	25.590
Network density	.001	.001
Number of clusters	18	10
Modularity value	.779	.463
Seeds' eigenvector centrality (EC)	[.0002; .0056]	[.0001; .0272]

Table 1: Key metrics of the two counter-message campaign networks

A modularity value of .779 indicates a high separation among the identified 18 clusters (resolution = 5)[5] of the JAK network, whereas .463 hints at a medium level of distinctiveness of the identified ten communities (resolution = 5) of the SMN campaign (Himelboim et al., 2013). The JAK seeds display very low values of EC, ranging from EC = .0002 to EC = .0056. Regarding the network SMN, the seeds show slightly higher but also very low values of EC, with four of five videos ranging from EC= .0001 to EC= .0087. Solely

one video displays a significantly higher value, namely EC = .0272. This video represents the most influential seed beneath the seeds that constitute the basis of the SMN (as well as the JAK) network. This seed also displays the highest number of views (n = 62,243) and the highest number of 248 incoming connections (also in comparison to all other nine seeds).

Network analysis: Jamal al-Khatib

The three largest clusters account for over 60% of all videos (see Figure 1 and Table 2). A brief characterization and the size of each cluster of the JAK network are shown in Table 2, whereas a more detailed description based on the content analysis sample can be found in Table 3.

Among the seven largest communities of the 18, four communities are dominated by PC: Community 10, 0, 15, and 9. Identified PC is rather subtle, as YouTube is legally bound to remove violent extremist and terrorist content. Community 10 is the second largest cluster (22.55%; see Table 2) and contains 80.7% IE propaganda (95% confidence interval = [0.78, 0.84]; see Table 3). Only 8.76% [0.06, 0.11] videos of Community 10 subsumed under the category Religion & Religious Music.Those videos mostly contain references to Islam but do not show a radical understanding of it. This cluster also contains other forms of PC, namely hate speech and RE propaganda, a few entertaining formats, and four (0.7% [0.00, 0.01]) CM (see Table 3). The fourth largest Community 0 (see Table 2) also includes a high number of IE propaganda amounting to 47.7% [0.42, 0.53] and a counterbalance of 42.1% [0.37, 0.48] of videos belonging to the category of either Music or Religion/Religious Music (Table 3). The Communities 15 and 9 are significantly smaller clusters than the first four or five (see Table 2) but are also dominated by PC, namely IE propaganda of both 58.1% [0.48, 0.69] (Table 3).

Two of the seed videos (B and D) belong to Community 10, which is the second largest (see Table 2) and most significant extremist cluster of the network (see Table 3). Another two seeds (A and E) belong to the largest cluster, Community 5 (see Table 2). This cluster consists almost exclusively of various entertaining categories except for CM (6.3% [0.04, 0.08]) as well as PC (3.8% [0.03, 0.05]; see Table 3). Each video category is represented in this cluster. The last seed (C) belongs to Community 13, which does not include either PC or CM (see Table 3): This cluster consists of videos assigned to the categories of People & Blogs (mainly lifestyle videos as YouTuber), Comedy (e.g., prank, sketch, parody) and Gaming (e.g., let's play, streaming). Most videos are People & Blogs (49% [0.42, 0.56], of which most show YouTubers reacting to other YouTubers' videos, often mockingly.

Figure 1: Size and description of the 18 communities and location of seeds (A-D) within the network JAK

Another cluster worth mentioning is the third largest Community 11 (see Table 2), which mainly consists of the category Music (70.7% [0.66, 0.75]; see Table 3) but also includes a few CM and PC. The eleven other communities are the smallest communities (see Table 2: grey and beige). They neither contain CM nor PC (see Table 3), and five comprise only one category (see Table 3). 759 of all 2,534 analyzed videos of the JAK network contain PC, which amounts to 30.0% [0.28, 0.32]. 44 of all 2,534 analyzed videos of the JAK network, namely 1.7% [0.01, 0.02], constitute CM other than the seeds.

In summary, in the case of the JAK campaign, for RQ1, the results show a strong connection between PC (most specifically Islamist extremist propaganda) and seed video B as well as D. In comparison to that, seeds A, C, and E show fewer relations to such content.

Community	Description	Size
5	Diverse Entertaining Content and Polluted Content	24.64%
10	IE Propaganda	22.55%
11	Music, Entertainment, and People/Blogs	14.13%
0	IE Propaganda, Religion/Religious Music, and Music	11.98%
13	People/Blogs, Comedy, and Gaming	7.59%
15	IE Propaganda, Religion/Religious Music, and Hate speech	3.40%
9	IE Propaganda, Religion/Religious Music, and Entertainment	3.38%
17	Religion/Religious Music and Entertainment	2.17%
1	Entertainment and Film	1.72%
16	Entertainment, Music, and News/Politics	1.60%
14	Animals, Entertainment, and Film	1.45%
2	Music (Arabic)	1.12%
12	Religion/Religious Music	.98%
3	Entertainment	.88%
4	Gaming	.82%
6	Music (German)	.82%
8	Education	.44%
7	Autos/Vehicles	.33%

Table 2: Legend of Figure 1: Size (percentage = share of videos/nodes in the network based on the entire dataset) and description (community label based on categories with highest shares of the content analysis) of the 18 communities within the network JAK

Regarding RQ2, the findings reveal that other CM than the seeds are underrepresented in the network, whereas 30.0% [0.28, 0.32] of all videos constitute PC. Community 5 has the highest CM share (6.3% [0.04, 0.08]). Besides this cluster, only two others of the 18 in total contain CM and their shares are below one percent. It can be constituted that other CM are per YouTube's definition of 'related videos' less associated with videos of the JAK campaign than videos containing PC. Regarding the titles of the five seed CM videos, it is suggested that keywords may have influenced the relatedness with PC (see also Schmitt et al., 2018): Whereas seed videos B and D, which belong to the biggest extremist cluster of the network, contain the words "shirk" and "takfir", the other three seed videos with fewer relations to PC do not include this explicit vocabulary that is often (mis-)used in Islamist extremist propaganda.

(in % [CI])	C0 (n=304)	C1 (n=44)	C2 (n=28)	C3 (n=22)	C4 (n=21)	C5 (n=624)	C6 (n=21)	C7 (n=8)	C8 (n=11)
Polluted content									
Conspiracy narratives						1.9 [0.01, 0.03]			
Hate speech						.2 [0.00, 0.00]			
RE Propaganda						.5 [0.00, 0.01]			
IE Propaganda	47.7 [0.42, 0.53]					1.3 [0.00, 0.02]			
Counter-messages									
Counter-messages						6.3 [0.04, 0.08]			
Other									
Autos & Vehicles						.2 [0.00, 0.00]		100.0 [1.00, 1.00]	
Comedy						1.1 [0.00, 0.02]			
Education						5.9 [0.04, 0.08]			100.0 [1.0, 1.0]
Entertainment	.3 [0.00, 0.01]	47.7 [0.33, 0.62]		95.5 [0.87, 1.04]		23.6 [0.20, 0.27]			
Film & Animation		36.4 [0.22, 0.51]		4.5 [−0.04, 0.13]		.8 [0.00, 0.02]			
Gaming					100 [1.0, 1.0]	.3 [0.00, 0.01]			
Howto & Style	.3 [0.00, 0.01]	2.3 [−0.02, 0.07]				3.2 [0.02, 0.05]			
Music	9.2 [0.06, 0.12]	13.6 [0.03, 0.24]	100.0 [1.0, 1.0]			1.3 [0.00, 0.02]	90.5 [0.78, 1.03]		
News & Politics	.3 [0.00, 0.01]					19.2 [0.16, 0.22]			
Nonprofits & Activism						1.4 [0.01, 0.02]			
People & Blogs						26.0 [0.23, 0.29]	4.8 [−0.04, 0.14]		
Pets & Animals							4.8 [−0.04, 0.14]		
Religion & Religious Music	42.1 [0.37, 0.48]					.8 [0.00, 0.02]			
Science & Technology						1.0 [0.00, 0.02]			
Sports						1.6 [0.01, 0.03]			
Travel & Events						3.5 [0.02, 0.05]			

Table 3: Overview of the composition of the 18 clusters in the network of JAK

(in % [CI])	C9 (n=86)	C10 (n=571)	C11 (n=358)	C12 (n=25)	C13 (n=192)	C14 (n=37)	C15 (n=86)	C16 (n=41)	C17 (n=55)
Polluted content									
Conspiracy narratives		$^{2}_{[0.00, 0.01]}$	$^{6}_{[0.00, 0.01]}$				$^{16.3}_{[0.08, 0.24]}$		
Hate speech		$^{1.4}_{[0.00, 0.02]}$	$^{1}_{[0.00, 0.02]}$						
RE Propaganda		$^{80.7}_{[0.78, 0.84]}$					$^{58.1}_{[0.48, 0.69]}$		
IE Propaganda	$^{58.1}_{[0.48, 0.69]}$								
Counter-messages									
Counter-messages		$^{7}_{[0.00, 0.01]}$	$^{3}_{[0.00, 0.01]}$						
Other									
Autos & Vehicles			$^{3}_{[0.00, 0.01]}$		$^{24.5}_{[0.18, 0.31]}$				
Comedy			$^{1.1}_{[0.00, 0.02]}$						
Education		$^{9}_{[0.00, 0.02]}$	$^{3}_{[0.00, 0.01]}$						
Entertainment	$^{14.0}_{[0.07, 0.21]}$	$^{9}_{[0.00, 0.02]}$	$^{14.8}_{[0.11, 0.18]}$		$^{4.2}_{[0.01, 0.07]}$	$^{27.0}_{[0.13, 0.41]}$		$^{56.1}_{[0.41, 0.71]}$	$^{32.7}_{[0.20, 0.45]}$
Film & Animation			$^{3}_{[0.00, 0.01]}$			$^{18.9}_{[0.06, 0.32]}$		$^{2.4}_{[-0.02, 0.07]}$	$^{1.8}_{[-0.02, 0.05]}$
Gaming		$^{4}_{[0.00, 0.01]}$			$^{18.8}_{[0.13, 0.24]}$				
Howto & Style		$^{1.4}_{[0.00, 0.02]}$			$^{1.6}_{[0.00, 0.03]}$				
Music		$^{2}_{[0.00, 0.01]}$	$^{70.7}_{[0.66, 0.75]}$		$^{.5}_{[0.00, 0.02]}$			$^{22.0}_{[0.09, 0.35]}$	$^{9.1}_{[0.01, 0.17]}$
News & Politics	$^{8.1}_{[0.02, 0.14]}$	$^{2.8}_{[0.01, 0.04]}$	$^{.3}_{[0.00, 0.01]}$					$^{19.5}_{[0.07, 0.32]}$	
Nonprofits & Activism					$^{49.0}_{[0.42, 0.56]}$	$^{8.1}_{[-0.01, 0.17]}$			
People & Blogs		$^{1.8}_{[0.01, 0.03]}$	$^{10.1}_{[0.07, 0.13]}$			$^{29.7}_{[0.15, 0.44]}$			
Pets & Animals									
Religion & Religious	$^{18.6}_{[0.10, 0.27]}$	$^{8.8}_{[0.06, 0.11]}$	$^{.3}_{[0.00, 0.01]}$	$^{100.0}_{[1.0, 1.0]}$	$^{.5}_{[0.00, 0.02]}$		$^{25.6}_{[0.16, 0.35]}$		$^{56.4}_{[0.43, 0.69]}$
Music						$^{5.4}_{[-0.02, 0.13]}$			
Science & Technology						$^{10.8}_{[0.01, 0.21]}$			
Sports	$^{1.2}_{[-0.01, 0.03]}$				$^{1.0}_{[0.00, 0.02]}$				
Travel & Events									

Table 3: Continued

Note: The table is based on samples of 15% we randomly drew for each cluster. Values are rounded. Confidence intervals are based on $\alpha = 5\%$.

Network analysis: Say My Name

The two largest communities account for over 90% of all collected videos (see Figure 2 and Table 4). A brief characterization of each cluster of the SMN network is shown in Table 4, whereas a more detailed analysis of each cluster's content can be found in Table 3.

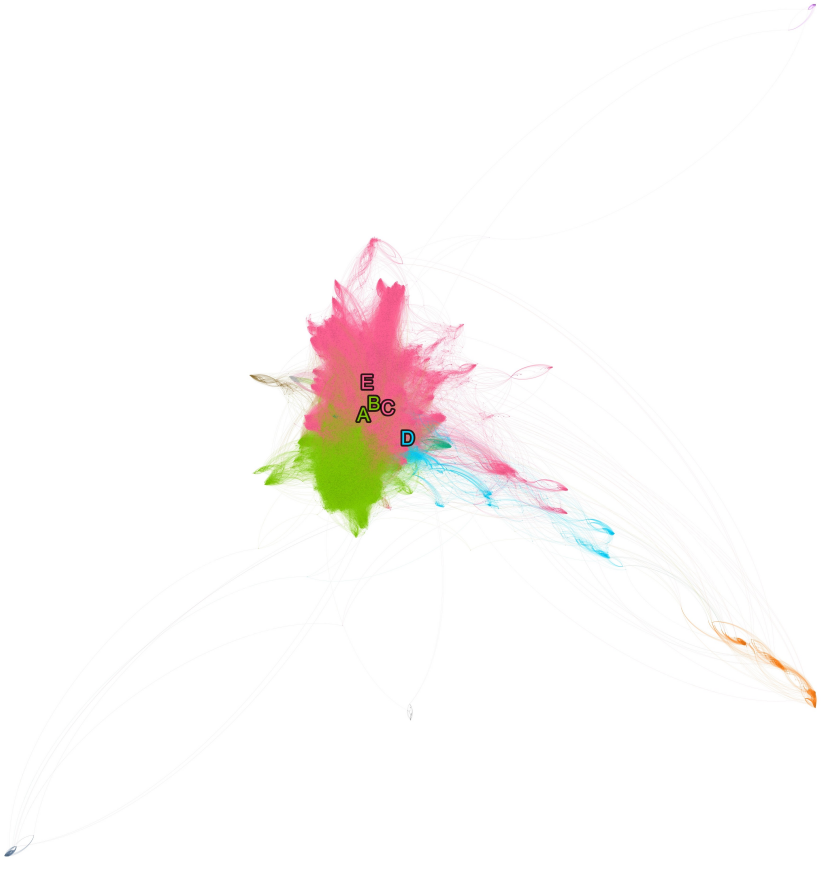

Figure 2: Size and description of the 18 communities and location of seeds (A-D) within the network JAK

The first two seed videos (A and B) are part of the second largest cluster (Community 1; see Figure 2 and Table 4): This cluster includes a wide range of diverse, entertaining content, whereby the category People & Blogs accounts for 71.9% [0.69, 0.75] and HowTo & Style for 16.0% [0.14, 0.18]. In this community, there are also other videos by the creator of the two seed

Community	Description	Size
9	Diverse Entertaining Content and Polluted Content	56.76%
1	People/Blogs and Diverse Entertaining Content	35.61%
0	Music and Howto/Style	3.03%
8	Music	2.01%
3	Howto and Gaming	.72%
2	Film and Entertainment	.45%
6	Entertainment, Travel, and People/Blogs	.42%
4	Sports and People/Blogs	.40%
5	Gaming	.39%
7	Howto	.20%

Table 4: Legend of Figure 2: Size (percentage = share of videos/nodes in the network based on the entire dataset) and description (community label based on categories with highest shares of the content analysis) of the ten communities within the network SMN

videos, which partially represent other CM. Regarding PC, solely two out of the 1,020 analyzed videos have been identified as hate speech (0.2% [0.00, 0.00]) and represent reactions of YouTubers to hate speech that contain hate speech themselves.

Another two of the five seed videos (C and E) belong to Community 9 (see Figure 2 and Table 4), constituting the network's largest cluster. A wide range of topics characterizes this community. Various entertaining videos are mainly categorized as People & Blogs and Entertainment. Further, various political videos are subsumed under the category News & Politics as general news/roundtable/political speech or People & Blogs as an individual's political opinion. This cluster has a high share of CM of .6% [0.05, 0.08]. Concerning PC, the results reveal that IE propaganda amounts to 6.6% [0.05, 0.08], RE propaganda to 1.4% [0.01, 0.02], hate speech to 0.1% [0.00, 0.00], and conspiracist videos to 1.1% [0.01, 0.02] (see Table 5). There were very few videos displaying Christian extremist beliefs and strong criticism of Islam, which were subsumed under hate speech due to their small amount but pejorative attitude. Besides this cluster and Community 1, which contains 0.2% [0.00, 0.00] hate speech, there is no other cluster with PC.

Table 5: Overview of the composition of the ten clusters in the network of SMN

(in % [CI])	C0 (n=87)	C1 (n=1020)	C2 (n=13)	C3 (n=21)	C4 (n=12)
Polluted content					
Conspiracy narratives					
Hate speech		0.2 [0.00, 0.00]			
RE Propaganda					
IE Propaganda					
Counter-messages					
Counter-messages	1.1 [−0.01, 0.03]	.7 [0.00, 0.01]			
Other					
Autos & Vehicles					
Comedy	4.6 [0.00, 0.09]	1.6 [0.01, 0.02]			
Education		1.0 [0.00, 0.02]			
Entertainment	1.1 [−0.01, 0.03]	5.0 [0.04, 0.06]	15.4 [−0.04, 0.35]		
Film & Animation		.1 [0.00, 0.00]	76.9 [0.54, 1.00]		
Gaming		.7 [0.00, 0.01]		23.8 [0.06, 0.42]	
Howto & Style	33.3 [0.23, 0.43]	16.0 [0.14, 0.18]		76.2 [0.58, 0.94]	8.3 [−0.07, 0.24]
Music	54.0 [0.44, 0.64]	.9 [0.00, 0.01]			
News & Politics					
Nonprofits & Activism		.1 [0.00, 0.00]			
People & Blogs	5.7 [0.01, 0.11]	71.9 [0.69, 0.75]			16.7 [−0.04, 0.38]
Pets & Animals		.2 [0.00, 0.00]			
Religion & Religious Music		.2 [0.00, 0.00]			
Science & Technology		.2 [0.00, 0.00]			
Sports		.1 [0.00, 0.00]	7.7 [−0.07, 0.22]		75.0 [0.51, 1.00]
Travel & Events		1.3 [0.01, 0.02]			

(in % [CI])	C5 (n=11)	C6 (n=12)	C7 (n=6)	C8 (n=58)	C9 (n=1626)
Polluted content					
Conspiracy narratives					1.1 [0.01, 0.02]
Hate speech					.1 [0.00, 0.00]
RE Propaganda					1.4 [0.01, 0.02]
IE Propaganda					6.6 [0.05, 0.08]
Counter-messages					
Counter-messages					6.6 [0.05, 0.08]
Other					
Autos & Vehicles					.2 [0.00, 0.00]
Comedy				1.7 [−0.02, 0.05]	2.9 [0.02, 0.04]
Education					8.7 [0.07, 0.10]
Entertainment		50.0 [0.22, 0.78]			19.1 [0.17, 0.21]
Film & Animation					1.1 [0.01, 0.02]
Gaming	100 [1.00, 1.00]				1.8 [0.01, 0.02]
Howto & Style			100.0 [1.0, 1.00]		4.7 [0.04, 0.06]
Music				96.6 [0.92, 1.01]	1.8 [0.01, 0.02]
News & Politics					11.2 [0.10, 0.13]
Nonprofits & Activism					1.7 [0.01, 0.02]
People & Blogs		16.7 [−0.04, 0.38]		1.7 [−0.02, 0.05]	21.7 [0.20, 0.24]
Pets & Animals					.2 [0.00, 0.00]
Religion & Religious Music					3.8 [0.03, 0.05]
Science & Technology					2.3 [0.02, 0.03]
Sports					2.3 [0.02, 0.03]
Travel & Events		33.3 [0.07, 0.60]			.6 [0.00, 0.01]

Table 5: Continued

Note: The table is based on samples of 15% we randomly drew for each cluster. Values are rounded. Confidence intervals are based on $\alpha = 5\%$.

The last seed (D) is part of the third largest community, 0 (see Figure 2 and Table 4). The cluster mainly contains videos of the category Music, namely English pop and German rap, and Howto & Style videos regarding hair, makeup, clothing, and recipes. The community also includes other videos of the creator of seed D, which partially represent other CM but contains no PC. All clusters other than Community 0, 1, and 9 within the network of SMN neither contain PC nor CM and are relatively small and coherent (Figure 2, Table 4 and 5). Community 9 represents the cluster with the highest share of both PC and CM and is also the largest cluster, with over 50% of all network videos. 151 of the 2,866 analyzed SMN videos constitute PC, representing 5.3% [0.05, 0.05] of all manually categorized videos. Other CM apart from the seeds of the network amount to 115, equaling 4.0% in total [0.04, 0.04].

In sum, regarding the campaign SMN compared to JAK, there are fewer associations between PC and CM. Answering RQ1, the findings reveal that, although there is only one cluster with more than two videos with PC and four out of five seeds' EC values are very low, there are connections between CM and PC. Although CM compete with a higher number of PC, they are directly related to the seeds and more widely spread across the clusters. Thus, regarding RQ2, it can be constituted, based on the scattering and proximity of CM, that the CM campaign SMN shows stronger associations to other CM than PC compared to the JAK campaign.

Discussion

This study combined network and content analysis to investigate how German CM campaigns on YouTube are connected to PC and other CM by YouTube's definition of 'related videos' that constitutes a fundamental aspect of YouTube's algorithmic recommendations. Results show that CM are not only connected to other CM but also closely connected to non-indictable PC, reinforcing the finding that "the exposure to CMs may be tainted with risks" (Schmitt et al., 2018, p. 801). Partially, PC is even directly connected to CM (CD = 1), which represents recommendations of videos containing PC based on prior watch of a video containing CM. Further, there remains a gap between the volume of CM and PC, which is particularly large regarding the JAK campaign. From these results, it can be concluded that the Network Enforcement Act may not have significantly impacted the subtle, non-indictable extremist propaganda. However, in comparison to the findings of Schmitt et al. (2018), the results display fewer (links to) conspiracist videos and hate speech. On the one hand, this could relate to the (different)

characteristics of the campaigns but is also likely to have been influenced by legal adjustments or YouTube's modification of its recommendation algorithm.

However, these results should not be interpreted as the success of attempts to stop the spread of PC online. On the contrary, the interrelatedness of CM and PC on YouTube due to the platform's algorithmic recommendations raises concerns. Speaking of differences between the two CM campaigns, findings indicate that the relatedness of CM and PC depends on both a) the campaign's design and b) its setup on YouTube: a) The SMN campaign targets a group that is thought to be potentially vulnerable to polluted online content and does not directly refer to extremist ideas or concepts but pursues a style of positive language, strengthening fundamental values in democracies, such as pluralism and tolerance. This campaign is associated with a comparatively low level of interrelatedness with polluted content (5.3% [0.05, 0.05] PC). The JAK campaign, on the contrary, aims at targeting individuals already sympathizing with jihadist or Salafist groups. It directly refers to extremist concepts and violence, also using respective words and visuals to transport messages against extremism and related forms. This campaign displays a high level of interrelatedness with polluted content (30.0% [0.28, 0.32] PC). b) Within the JAK campaign, results further reveal that keywords, especially regarding the videos' titles, may be a crucial factor driving the interrelatedness with polluted content (seed videos B and D mention "shirk" and "takfir" and belong to the network's biggest extremist cluster).

Portraying CM as normatively "good" therefore needs to be challenged for three reasons: First, the structural proximities to PC pose severe threats, as internal referrals on YouTube have been found to be the main factor leading to exposure to videos besides the active search (Zhou et al., 2010). Second, regarding the potential effects of unintentional exposure due to recommendations, content labeled as 'related' may encourage users to concern themselves with content from attitude-inconsistent sources (Messing & Westwood, 2014), potentially resulting in the acceptance of unexpected messages. Third, online propaganda seems to be especially capable of confirming pre-existing extremist beliefs (Wojcieszak, 2009). However, CM can ideal-typically be theorized as (positive) messages against polluted content. Referring to our theoretical framework, we suggest differentiating between intentionality and truthfulness (Wardle, 2018) to generate less normative evaluations of these types of content.

Limitations and future perspectives

Several limitations to this research are pointed out to provide improvements and ideas for future studies on the interrelatedness of PC and CM. First, the empirical study is limited to two specific CM campaigns that are not representative of all. Second, the two information network analyses have been conducted based on data gathered at one point in time and thus do not allow a longitudinal view. Third, data has only been gathered with a crawl depth of two. Addressing these three limitations, future research may profit from collecting data at different times (Courtois & Timmermans, 2018) and multiple campaigns with an increased crawl depth to generalize results for the interrelatedness of CM and PC on YouTube. Fourth, the manual content analysis was only conducted on a 15% sample of each cluster of each network. Addressing the first three limitations by scaling up the research intensifies the difficulty of conducting a reliable and valid content analysis of the entire dataset: The identification of polluted content is challenging, as extremist actors seek to subtly communicate their radical ideas to bypass the prohibition and deletion of content relevant to criminal law. Future research will benefit greatly from further development of automated content analysis. Fifth, this research strongly depends on what YouTube provides via the data API (Google Developers, 2017), e.g., limiting the maximum number of "related videos". Sixth, data has been gathered with the user profile of a "blank prototype", meaning that browser and download history, cookies and other website data have been deleted. A simulation of different user types, as Schmitt et al. (2018) suggested, or data donations would further increase the external validity.

Practical implications

From a practical lens, the results of this study provide empirically sound starting points for the design of CM for prevention actors: It shows that the dissemination of positive counter-message campaigns (e.g., SMN) via YouTube is associated with a comparatively low level of interrelatedness with PC. Simultaneously, the results do not imply that CM campaigns directly referring to extremist ideas (e.g., JAK) should not be utilized. It is vital to consider the campaign's design in conjunction with the type of prevention and target group: The JAK campaign displays close connections to PC on YouTube but holds the potential of reaching already radicalized individuals and displays a narrative, two-sided style (male protagonist's emotional stories of exiting extremist groups) that might hold the potential to avoid

reactance effects when being consumed (Schmitt et al., 2021). Another factor represents the distribution and setup via YouTube: On the one hand, public dissemination via YouTube of prevention campaigns that contain direct references to polluted content must be questioned critically: Is there a risk that (non-radicalized) individuals will be exposed to the videos and in turn potentially directed to polluted content? On the other hand, there is already more published polluted content than counter-messages on YouTube, and the connectedness of CM and PC might also offer the chance of breaking up a recommendation spiral of videos with PC. However, a further 'meta-analysis' of the underlying dataset of the two CM networks shows that the relatedness of CM and PC is asymmetrical in a way that there is more CM 'recommending' PC than PC relate or link to CM (Zieringer, 2022), which in parts can be again traced back to the higher amount of PC. Ultimately, a publication of CM campaigns, and especially those for already radicalized individuals, on YouTube serves the objective of working against the vast amount of PC on the platform but is recommended to be set up with a careful choice of keywords that contain the positively framed message of the videos with less direct references to extremist concepts.

Beyond these prevention measures, the promotion of media literacy as well as community management and moderation, play an outstanding role: Recognition of Schmitt et al. that "online and offline [countering and preventing violent extremism] efforts should be combined in order to successfully counter the negative effects of [polluted content]" (2018, p. 801), is still relevant today. However, responsibility should not only be shifted to media literacy education and prevention actors: The research provides insight into a significant effort to find evidence regarding ideological biases in algorithmic recommendations. The conclusion of current research, that "as the nature of opinion power [including the systemic power of platforms] is changing, so must the tools of control" (Seipp et al., 2023, p. 1) is again underscored (for Germany: Reinemann & Zieringer, 2021; Stegmann et al., 2022). Regarding current policy initiatives in Europe, the imposition of greater societal responsibility on media platforms runs the risk of amplifying the platforms' power as "active political actors in their own right" (Helberger, 2020, p. 842). Design principles for recommender systems represent one solution (see Helberger et al., 2018). The primary goal of exposure diversity should be interpreted based on democratic theory within its bounds concerning the potential interrelatedness of ideologically diverse counter-messages and polluted content on YouTube.

Notes

[1]LGBT+ is an acronym for lesbian, gay, bisexual, and transgender and includes other forms of gender and sexual identification or orientation, i.e., queer, intersex, and asexual.

[2]Takfir means, according to Islamic law and theology, the ex-communication of Muslims declared as unbelievers, i.e., Kāfir (Hassan, 2017).

[3]Shirk is a strongly negative connotated term for associating other entities with God (Esposito, 2009).

[4]For more information on the two campaigns and the seed videos' descriptive variables, please see Online Appendix A and B: https://osf.io/vybg3/

[5]The resolution of five has been selected based on the approach to get as distinct clusters as possible and as many as necessary. Schmitt and colleagues have used the same resolution for their study (2018).

Data Availability Statement

The data underlying this article are available in the article and its online supplementary material, which can be found at: https://osf.io/vybg3/ The underlying raw data of this article will be shared on reasonable request to the corresponding author.

References

Al-Taie, M. Z., & Kadry, S. (2017). *Python for Graph and Network Analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-319-53004-8

Arnold, K. (2003). Propaganda als ideologische Kommunikation. [Propaganda as ideological communication]. *Publizistik*, *48*(1), 63–82.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi. The open graph viz platform. Retrieved June 3, 2023, from https://gephi.org/

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bode, L., & Vraga, E. K. (2015). In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication*, *65*(4), 619–638. https://doi.org/10.1111/jcom.12166

Braddock, K. (2022). Vaccinating Against Hate: Using Attitudinal Inoculation to Confer Resistance to Persuasion by Extremist Propaganda. *Terrorism and Political Violence*, *34*(2), 240–262. https://doi.org/10.1080/09546553.2019.1693370

Caplan, G., & Caplan, R. B. (2000). The Future of Primary Prevention. *The Journal of Primary Prevention*, *21*, 131–136. https://doi.org/10.1023/A:1007062631504

Carter, E. (2018). Right-wing extremism/radicalism: Reconstructing the concept. *Journal of Political Ideologies*, *23*(2), 157–182. https://doi.org/10.1080/13569317.2018.1451227

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior*, *63*, 311–320. https://doi.org/10.1016/j.chb.2016.05.033

Courtois, C., & Timmermans, E. (2018). Cracking the Tinder Code: An Experience Sampling Approach to the Dynamics and Impact of Platform Governing Algorithms. *Journal of Computer-Mediated Communication*, *23*(1), 1–16. https://doi.org/10.1093/jcmc/zmx001

Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198. https://doi.org/10.1145/2959100.2959190

Dahlgren, P. M. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, *42*(1), 15–33. https://doi.org/10.2478/nor-2021-0002

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., & Sampath, D. (2010). The YouTube video recommendation system. *Proceedings of the fourth ACM conference on Recommender systems*, 293–296. https://doi.org/10.1145/1864708.1864770

Esposito, J. L. (2009). *The Oxford Encyclopedia of the Islamic World*. Oxford University Press. https://doi.org/10.1093/acref/9780195305135.001.0001

Faddoul, M., Chaslot, G., & Farid, H. (2020). A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. *arXiv*, *2003.03318*. Retrieved June 3, 2023, from http://arxiv.org/abs/2003.03318

Figueiredo, F., Benevenuto, F., & Almeida, J. M. (2011). The tube over time: Characterizing popularity growth of youtube videos. *Proceedings of the fourth ACM international conference on Web search and data mining*, 745–754. https://doi.org/10.1145/1935826.1935925

Filippova, K., & Hall, K. B. (2011). Improved video categorization from text metadata and user comments. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 835–842. https://doi.org/10.1145/2009916.2010028

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. https://doi.org/10.1093/poq/nfw006

Frischlich, L., Rieger, D., Morten, A., & Bente, G. (2018). The Power of a Good Story: Narrative Persuasion in Extremist Propaganda and Videos against Violent Extremism. *International Journal of Conflict and Violence (IJCV)*, *12*, 1–16. https://doi.org/10.4119/UNIBI/IJCV.644

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000233231

Google Developers. (2017). Search: List | YouTube Data API. Retrieved June 3, 2023, from https://developers.google.com/youtube/v3/docs/search/list

Gottfried, J., & Shearer, E. (2016). News use across social media platforms 2016. *Pew Research Center*. Retrieved June 3, 2023, from https://www.journalism.org/wp-content/uploads/sites/8/2016/05/PJ2016.05.26social-media-and-newsFINAL-1.pdf

Gwet, K. L. (2008). Intrarater Reliability. In D'Agostino, R.B. and Sullivan, L. and Massaroa, J. (Ed.), *Wiley Encyclopedia of Clinical Trials*. John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9780471462422.eoct631

Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z., & Wojcieszak, M. (2022). YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations. *arXiv*, *2203.10666*. Retrieved June 3, 2023, from http://arxiv.org/abs/2203.10666

Hassan, M. H. (2017). The Danger of Takfir (Excommunication): Exposing IS' Takfiri Ideology. *Counter Terrorist Trends and Analyses*, *9*(4), 3–12. https://www.jstor.org/stable/26351508

Helberger, N. (2020). The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power. *Digital Journalism*, *8*(6), 842–854. https://doi.org/10.1080/21670811.2020.1773888

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

Hemmingsen, A.-S., & Castro, K. I. (2017). *The trouble with counter-narratives*. Danish Institute for International Studies. http://pure.diis.dk/ws/files/784884/DIISRP20171.pdf

Himelboim, I., Smith, M., & Shneiderman, B. (2013). Tweeting Apart: Applying Network Analysis to Detect Selective Exposure Clusters in Twitter. *Communication Methods and Measures*, *7*(3-4), 195–223. https://doi.org/10.1080/19312458.2013.813922

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, *9*(6), e98679. https://doi.org/10.1371/journal.pone.0098679

Jamal al-Khatib. (2020). YouTube. Channel info. Jamal al-Khatib. Retrieved June 3, 2023, from https://www.youtube.com/channel/UCKmWuKvMLGHQ4Z0VaVjwYVQ/about

Jürgens, P., & Stark, B. (2022). Mapping Exposure Diversity: The Divergent Effects of Algorithmic Curation on News Consumption. *Journal of Communication*, *72*(3), 322–344. https://doi.org/10.1093/joc/jqac009

Keeley, B. L. (1999). Of Conspiracy Theories. *The Journal of Philosophy*, *96*(3), 109–126. http://www.jstor.org/stable/2564659

Knudsen, E. (2023). Modeling news recommender systems' conditional effects on selective exposure: Evidence from two online experiments. *Journal of Communication*, *73*(2), 138–149. https://doi.org/10.1093/joc/jqac047

Krafft, T. D., Gamer, M., & Zweig, K. A. (2019). What did you see? A study to measure personalization in Google's search engine. *EPJ Data Science*, *8*(38). https://doi.org/10.1140/epjds/s13688-019-0217-5

Landesanstalt für Medien NRW [State Media Authority North Rhine Westphalia]. (2022). Hate Speech forsa-Studie 2022. Zentrale Untersuchungsergebnisse. [Hate Speech forsa study 2022. Central study results]. https://www.medienanstalt-nrw.de/themen/hass/forsa-befragung-zur-wahrnehmung-von-hassrede.html

Ledwich, M., & Zaitsev, A. (2019). Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization. *arXiv*, *1912.11211*. Retrieved June 3, 2023, from http://arxiv.org/abs/1912.11211

Loecherbach, F., Moeller, J., Trilling, D., & van Atteveldt, W. (2020). The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism*, *8*(5), 605–642. https://doi.org/10.1080/21670811.2020.1764374

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Mattis, N., Masur, P., Möller, J., & van Atteveldt, W. (2022). Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommender design. *New Media & Society*. https://doi.org/10.1177/14614448221104413

Messing, S., & Westwood, S. J. (2014). Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research*, *41*(8), 1042–1063. https://doi.org/10.1177/0093650212466406

Morris, E. (2016). Children: Extremism and online radicalization. *Journal of Children and Media*, *10*(4), 508–514. https://doi.org/10.1080/17482798.2016.1234736

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(026113), 1–15. https://doi.org/10.1103/PhysRevE.69.026113

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Reuters Institute Digital News Report 2022*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/DigitalNews-Report2022.pdf

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. *Proceedings of the 23rd international conference on World wide web*, 677–686. https://doi.org/10.1145/2566486.2568012

Nienierza, A., Reinemann, C., Fawzi, N., Riesmeyer, C., & Neumann, K. (2019). Too dark to see? Explaining adolescents' contact with online extremism and their ability to recognize it. *Information, Communication & Society*, *24*(9), 1229–1246. https://doi.org/10.1080/1369118X.2019.1697339

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (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

Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You*. Penguin.

Rainie, L., & Anderson, J. (2017). Code-dependent: Pros and cons of the algorithm age. *Pew Research Center*. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2017/02/PI2017.02.08AlgorithmsFINAL.pdf

Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., Oksanen, A., Räsänen, P., & Zych, I. (2021). Hate Knows No Boundaries: Online Hate in Six Nations. *Deviant Behavior*, *42*(9), 1100–1111. https://doi.org/10.1080/01639625.2020.1722337

Reinemann, C., Nienierza, A., Fawzi, N., Riesmeyer, C., & Neumann, K. (2019). *Jugend - Medien - Extremismus: Wo Jugendliche mit Extremismus in Kontakt kommen und wie sie ihn erkennen [Youth - media - extremism: Where adolescents come into contact with extremism and how they recognize it]*. Springer VS. https://doi.org/10.1007/978-3-658-23729-5

Reinemann, C., & Zieringer, L. (2021). *Meinungsmachtkontrolle und Vielfaltsmonitoring im digitalen Zeitalter: Eine kritische Reflexion der Begriffe, Annahmen, Indikatoren und Verfahren von Medienstaatsvertrag, Konzentrationskontrolle und Medienvielfaltsmonitoring [Controlling opinion power and monitoring diversity. A critical reflection on the terms, assumptions, indicators, and procedures of the state media treaty, concentration control, and media diversity monitor]*. Bavarian Research Institute for Digital Transformation (bidt) [Working Paper No. 4]. https://doi.org/10.35067/BV16-2Z30

Rieder, B. (2015). YouTube Data Tools video network. Retrieved June 3, 2023, from https://tools.digitalmethods.net/netvizz/youtube/modvideosnet.php

Rieger, D., Frischlich, L., & Bente, G. (2013). *Propaganda 2.0: Psychological effects of right-wing and islamic extremist internet videos*. Luchterhand.

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (2021). Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit. *Social Media + Society*, *7*(4), 1–14. https://epub.ub.uni-muenchen.de/92794/1/20563051211052906.pdf

Rieger, D., Schmitt, J. B., & Frischlich, L. (2018). Hate and counter-voices in the Internet: Introduction to the special issue. *Studies in Communication | Media*, *7*(4), 459–472. https://doi.org/10.5771/2192-4007-2018-4-459

Röchert, D., Neubaum, G., Ross, B., Brachten, F., & Stieglitz, S. (2020). Opinion-based Homogeneity on YouTube. *Computational Communication Research*, *2*(1), 81–108. https://computationalcommunication.org/ccr/article/view/15/7

Sängerlaub, A., & Schulz, L. (2021). Desinformation in Sozialen Medien [Desinformation in social media]. *Reset. & pollytix*. https://public.reset.tech/documents/210811ResetpollytixDesinformation.pdf

Santos-d'Amorim, K., & Miranda, M. K. F. O. (2021). Misinformation, disinformation, and malinformation: Clarifying the definitions and examples in disinfodemic times. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, *26*, 01–23. https://doi.org/10.5007/1518-2924.2021.e76900

Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: Options and limitations. *Digital Policy, Regulation and Governance*, *17*(6), 35–49. https://doi.org/10.1108/info-05-2015-0025

Say my Name. (2020). YouTube. Channel info. Say My Name. Retrieved June 3, 2023, from https://www.youtube.com/channel/UC4-UEgR6PFHfeFLvtzwTqww/about

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*. https://doi.org/10.1177/14614448221091185

Schmitt, J. B., Caspari, C., Wulf, T., Bloch, C., & Rieger, D. (2021). Two sides of the same coin? The persuasiveness of one-sided vs. two-sided narratives in the context of radicalization prevention. *Studies in Communication and Media*, *10*(1), 48–71. https://doi.org/10.5771/2192-4007-2021-1-48

Schmitt, J. B., Ernst, J., Rieger, D., & Roth, H.-J. (Eds.). (2020). *Propaganda und Prävention [Propaganda and Prevention]*. Springer VS. https://doi.org/10.1007/978-3-658-28538-8

Schmitt, J. B., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube. *Journal of Communication*, *68*(4), 780–808. https://doi.org/10.1093/joc/jqy029

Schulze, H., Hohner, J., & Rieger, D. (2022). Soziale Medien und Radikalisierung [Social Media and Radicalization]. In L. Rothenberger, J. Krause, J. Jost, & K. Frankenthal (Eds.), *Terrorismusforschung - Interdisziplinäres Handbuch für Wissenschaft und Praxis [Handbook Terrorism Research]* (pp. 319–330). Nomos. https://doi.org/10.5771/9783748904212-319

Seipp, T. J., Helberger, N., de Vreese, C., & Ausloos, J. (2023). Dealing with Opinion Power in the Platform World: Why We Really Have to Rethink Media Concentration Law. *Digital Journalism*, 1–26. https://doi.org/10.1080/21670811.2022.2161924

Sonck, N., Livingstone, S., Kuiper, E., & de Haan, J. (2011). *Digital literacy and safety skills*. EU Kids Online, London School of Economics & Political Science. http://eprints.lse.ac.uk/33733/

Stegmann, D., Zieringer, L., Stark, B., & Reinemann, C. (2022). Meinungsvielfalt, Meinungsmacht, Meinungsbildung. Zum (ungeklärten) Verhältnis zentraler Begriffe der deutschen Medienkonzentrationskontrolle [Diversity of opinion, opinion power, opinion formation - On the (unresolved) relationship of central concepts of German media concentration control]. *UFITA*, *86*(1), 38–70. https://doi.org/10.5771/2568-9185-2022-1-38

Stroud, N. J. (2010). Polarization and Partisan Selective Exposure. *Journal of Communication*, *60*(3), 556–576. https://doi.org/10.1111/j.1460-2466.2010.01497.x

Thorson, K., Cotter, K., Medeiros, M., & Pak, C. (2021). Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society*, *24*(2), 183–200. https://doi.org/10.1080/1369118X.2019.1642934

Wardle, C. (2018). The Need for Smarter Definitions and Practical, Timely Empirical Research on Information Disorder. *Digital Journalism*, *6*(8), 951–963. https://doi.org/10.1080/21670811.2018.1502047

Wojcieszak, M. (2009). "Carrying Online Participation Offline"-Mobilization by Radical Online Groups and Politically Dissimilar Offline Ties. *Journal of Communication*, *59*(3), 564–586. https://doi.org/10.1111/j.1460-2466.2009.01436.x

YouTube. (2019). Continuing our work to improve recommendations on YouTube. *blog.youtube*. Retrieved June 3, 2023, from https://blog.youtube/news-and-events/continuing-our-work-to-improve/

Zhou, R., Khemmarat, S., & Gao, L. (2010). The impact of YouTube recommendation system on video views. *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 404–410. Retrieved June 3, 2023, from https://dl.acm.org/doi/10.1145/1879141.1879193

Zieringer, L. (2022). YouTube recommendation algorithms' potential role in suggesting polluted content based on prior watch of counter-messages – A meta-analysis. *Book of Abstracts of ECREA's 9th European Communication Conference*, 252–253. Retrieved June 3, 2023, from https://conferences.au.dk/fileadmin/conferences/2022/ECREA/ECREA2022-AbstractBook.pdf

Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., de 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