ABSTRACT

Around 2018, YouTube became heavily criticized for its radicalizing function by allowing far-right actors to produce hateful videos that were in turn amplified through algorithmic recommendations. Against this 'algorithmic radicalization' hypothesis, Munger and Phillips (2019) argued that far-right radical content on YouTube fed into audience demand, suggesting researchers adopt a 'supply and demand' framework. Navigating this debate, our article deploys novel methods for examining radicalization in the language of far-right pundits and their audiences within YouTube's so-called 'Alternative Influence Network' (Lewis, 2018). To that end, we operationalize the concept 'extreme speech'—developed to account for 'the inherent ambiguity of speech contexts' online (Pohjonen and Udupa, 2017)—to an analysis of a right-wing 'Bloodsports' debate subculture that thrived on the platform at the time. Highlighting the topic of 'race realism', we develop a novel mixed-methods approach: repurposing the far-right website Metapedia as a corpus to detect unique terms related to the issue. We use this corpus to analyze the transcripts and comments from an archive of 950 right-wing channels, collected from 2008 until 2018. In line with Munger and Phillips' framework, our empirical study identifies a market for extreme speech on the platform, which came into public view in 2017.

1 Introduction

In January 2017, YouTube launched their 'Super Chat' feature, a new way to 'monetise your channel through the YouTube partner Programme' and 'build your community quickly' (Google, 2017). The feature was taken up enthusiastically by a variety of 'niche entrepreneurs' (Rieder *et al.*, 2018, p. 64), and a year after its introduction saw one YouTube channel bring in several thousand dollars during a livestream debate that attracted 10,000 viewers and became the top-trending live video worldwide (Lewis, 2018, p. 3). The topic was white nationalism. Alongside various YouTube microcelebrity 'skeptics' and contrarians, Richard Spencer—the 'dapper' founder of altright.com—spent 4 h with Carl Benjamin, going by the name Sargon of Akkad, extolling the virtues of the white ethnostate, while channel hosts and fellow debaters giggled about racism and studiously discussed the inherent biological superiority of whiteness. Marketed as 'Internet Bloodsports', the debate exemplified several of the factors that had driven the growth and visibility of the 'alt-right', notably the discursive alignment of self-styled contrarian 'classical' liberals (read libertarian) and the far-right, and the strategic negotiation of platform affordances for ideological gain.

In fact, in 2018, of all the prominent Silicon Valley platforms, it was YouTube that would be critiqued for providing a 'safe space' for far-right influencers and their ideas (Dunphy, 2017; Hern, 2018). Analysis of online political subculture (Lewis, 2018) revealed how these far-right influencers were partly successful in disseminating their content within a broader 'Alternative Influence Network' (AIN). While the political views of influencers in Lewis' AIN range 'from mainstream versions of libertarianism and conservatism, all the way to overt white nationalism', they may all generally be characterized as adopting a 'contrarian' stance toward progressive ideals and a demand for 'dangerous' ideas to be constantly entertained (Lewis, 2018, p. 1). This contrarian stance might help explain their apparent eagerness to engage with issues portrayed as 'too controversial' to receive a fair hearing in 'mainstream' intellectual debate (Lewis, 2018; Weiss, 2018).

While Lewis' Data and Society report described a vast online (sub)cultural phenomena, a 'moral panic' (Livingstone, 2019) ensued concerning YouTube's mediation of political engagement and especially the ways its recommendation engine served as a 'radicalization machine' (Tufekci, 2018), funneling users toward more extreme political content (Tokmetzis, 2019). Kevin Roose, a journalist at The New York Times, wrote an article and recorded an entire podcast series about Caleb Cain's story, a young YouTuber who told the reporter how he 'fell down the alt-right rabbit hole' (Roose, 2019). While Caleb Cain's story presents a complex mixture of teenage angst, social alienation, and alternative (online) media consumption, Roose posits that when observing the stories of those that went down the far-right rabbit hole, the 'common thread in many of these stories is YouTube and its recommendation algorithm' (Roose, 2019).

In 2019, several working papers were released that applied quantitative techniques to relatively large-scale datasets to measure the recommendation algorithm's impact on the growth of right-wing radicalization on the platform (Ribeiro *et al.*, 2019; Ledwich and Zaitsev, 2019).[2] While

there was significant disagreement between Ribeiro *et al.* and Ledwich and Zaitsev about the 'political bias' of the recommendation system, the outcome suggested that large-scale mapping of algorithms via YouTube's API can never really capture the 'actual experience' of the audience (Feuer, 2019). In addition to the methodological issues, the political scientists Munger and Phillips (2019, 2020) argued that it was not exclusively (or even primarily) YouTube's recommendation engine that was the source of radicalization, but that a demand for radical content among audiences already existed, pointing to the high levels of community engagement between users and the ideas expressed in videos by far-right pundits. Rather than 'radicalization by algorithm', they argued that 'the true threat posed by YouTube is the capacity to create '*radical alternative political canons* and interpretive communities to match' (2019, p. 6, emphasis added). Their argument then presents somewhat of a return to Lewis' (2018) earlier observations about the marketing dynamics of YouTube radicalization and how influencers can radicalize their audiences, each other, or even be '*radicalized by their own audience's engagement*' (Lewis, 2018, p. 6, emphasis added).

This article builds on Munger and Phillips' basic line of argument about active audience participation and Lewis' observation of YouTube's subculture of 'race realism Bloodsports'. However, instead of quantifying recommendations or influencer 'vanity' metrics, we focus primarily on examining the role of transgressive 'extreme speech'—a concept that seeks to highlight the contextual ambiguity and 'textured nature of online abuse' (Pohjonen and Udupa 2017, p. 1174)—in videos and the comment sections engaging with one of the central issues fueling platform controversy: race (Lewis, 2018). In doing so, we follow a particular suggestion—made by Ribeiro *et al.* (2019)—that future quantitative research concerning user radicalization on YouTube might 'trace the evolution of the speech of content creators and commenting users throughout the years' (p. 10). By combining digital methods and natural language processing techniques for analyzing hate speech, this article offers an empirically grounded and contextualized account of a significant moment when extremists plagued the platform and tested Silicon Valley's 'free speech' ideals. Our methodology may offer digital humanities research ideas for how to study the complex dynamics that tend to underlie the use of extreme speech online, which includes multiple 'actors' each possessing a degree of 'agency'—from the platform, to the microcelebrity, to the audience right down to the vernacular.

Following Ribeiro *et al.*, we endeavor to map the evolution of 'extreme speech' practices across video transcripts and comments of 950 right-wing channel videos collected by journalist and extremism expert Dimitri Tokmetzis (see Tokmetzis, 2019). We focus on one specific type of extreme speech: pseudo-scientific debates around 'race realism', which may also be understood as an 'intellectualized' type of discriminatory discourse in distinction to hate speech slang. Concluding our distant reading of transcripts and comments with an exemplary case study drawn from the so-called 'Bloodsports' debate culture that proliferated on YouTube in 2017–18, we present the rise of 'race realism' on the platform within the context of historically contingent factors, including the rhetorical negotiation between content creators and platform regulations. Methodologically, our thesis is that in order to detect and examine the dynamics of 'extreme speech', one can focus on the broader evolution of speech tied to the norms it seeks to transgress. That is, instead of relying on language that is already radicalized—for example, a predetermined list of hate speech terms—we examine the evolution of speech related to 'race' as a concept widely uttered by online right-wing political subcultures to debate, determine, and transgress norms of racial tolerance and identity. From there, we are able to comment on a variety of discourses that break such norms, including the development of problematic language that does not formally constitute hate speech slang but is discriminatory in substance, like scientific racism. In this sense, we conclude that hate speech detection may constitute less a detection of specific hateful terms than detecting the contingent vocabulary of a variety of discriminatory ideologies.

We find that the language of scientific racism, employed by far-right actors discussing race, fits within a particular 'rational' debating culture that was popular among YouTube's alternative influencers and their audience at the time of this study. This type of discourse emerged in the context of wider debates about how to speak about race and racism, and went initially undetected by content moderators designed to capture hate speech slang. It may then be argued that this 'intellectualized' type of hate speech was a byproduct of an attempt by YouTube's alternative influencers at promoting their own microcelebrity in a relatively unmoderated environment of political discussion, which permitted the dissemination of 'extreme speech' in the name of free speech. In addition to the much-discussed role of YouTube's algorithms, we conclude that efforts of far-right political figures to find new audiences and channel hosts' willingness to 'platform' controversial debate contributed to a growing audience demand and engagement with extreme political content on YouTube.

2 Theoretical Framework

Long before the 'fake news debacle' (Bounegru *et al.*, 2017)—long even before the 'filter bubble' hypothesis (Pariser, 2011)—scholarship had noted the web's tendency to encourage homophila 'not in demographic terms, but in terms of interest and outlook' within 'balkanized speech markets' (Sunstein, 2001, p. 69). Half a decade ago, social media platforms such as Google and Facebook promoted the notion of 'connectivity' as an unalloyed social good that 'benefits everyone', even as a 'human right' (Schmidt and Cohen, 2014; Zuckerberg, 2013). However, over the last few years, such assessments have been radically revised, with observers noting that social media connectivity can also strengthen 'communities . . . bound by hatred and prejudice' (Zuckerberg, 2018, p. 3).

Over the last years, YouTube has especially come under heavy criticism for hosting hateful communities as part of the platform's tolerance for free speech (Hokka, 2021). To this day, YouTube's attempt to 'fix' hate speech appears problematic, as critics continue to stress how YouTube's policy on hate speech 'remains vague' (Stokel-Walker, 2019). While some of this vagueness can be attributed to YouTube's neoliberal approach, it also has to do with the inherent ambiguity of user-generated content and especially the transgressive and highly ironic type of language thriving on YouTube (Hokka, 2021). Pohjonen and Udupa (2017), who argue that the concept of hate speech is often ill-suited to empirically based media and communications research, suggest employing the concept of 'extreme speech' as an alternative analytical concept. As Pohjonen and Udupa (2017) define it, the concept of 'extreme speech' acknowledges 'the inherent ambiguity of speech contexts' in online communication, where actors seek to 'push the boundaries of acceptable norms of public culture toward what the mainstream considers a breach within historically constituted normative orders' (p. 1174). To grasp how and why actors engage in extreme speech, they recommend attending to 'different situational features, including technology, online agency, and political cultures' (ibid, p. 1176).

In a pioneering qualitative report on YouTube's political debate culture, communications scholar Becca Lewis (2018) identified an AIN of political YouTube pundits who gained increased prominence as a provocative debating subculture in the late 2010s. In this report and a more recent academic article (Lewis, 2019), she claims that many of these pundits can be examined in relation to subcultural rules that govern microcelebrities on the platform. Despite the 'serious' political ambitions of some figures in the AIN, their debating culture should also be understood as a product of a broader entertainment marketplace. The entertainment here is a type of debate often characterized by a particular type of logical pedantry, popular with a young and technologically literate male demographic attuned to vernacular message boards like Reddit and 4chan—both milieux whose 'Internet trolling' subculture has been associated with the rise of distinctly alternative styles of reactionary political speech and punditry (Massanari, 2017; Nagle, 2017; Beran, 2019).

Lewis (2019) notes that successful YouTube microcelebrities tend to 'build trust with their audiences by stressing their relatability, their authenticity, and their accountability to those

audiences' (p. 4). Lewis (2018) thereby documents how pundits across a relatively broad spectrum of political thought appear on each other's YouTube channels in one-on-one discussions, or occasionally as part of larger 'Bloodsports' debates orchestrated in public livestreams. Through these events, they seek to expand their microcelebrity status by demonstrating debating acumen about particularly controversial topics. 'Bridges' or intermediates in the network introduce and translate more extreme ideas to broader audiences. One figure Lewis offers as an 'intermediary' is Carl Benjamin ('Sargon of Akkad'), who debates avowedly white nationalist 'alt-right' figures while cultivating a libertarian persona unafraid of dangerous ideas. As such, Lewis argues that the AIN served as a 'cross-promotion network' where radical 'white nationalist' ideas became relatively accessible to a vast audience of political pundits in the network (e.g. Jordan Peterson's 3.6 million subscribers to date).

Although Lewis claims that this 'cross-promotion of ideas forms a broader "reactionary" position' (2018, p. 1), attaching any single descriptor to this 'alternative' culture of political debate is challenging.[3] One value that seems however to be shared across the AIN is a firm belief in 'free speech'. In reference to European charters of human rights and the First Amendment of the American Constitution, pundits and audiences alike decry any attempted platform censorship, framing it as a violation of their innate right to speak and safeguard ideological diversity. As has been widely noted, a robust defense of speech 'freed' from the perceived normative constraints of postwar liberal political values (tolerance, equality, anti-discrimination) has become one of the defining issues in efforts to 'unite the right' (Davey and Ebner, 2017; Weisman, 2018; Wendling, 2018; Hermansson *et al.*, 2020). In the context of 'alternative' political discussion online, the notion of free speech on YouTube may also be understood in relation to older 'hacker' subcultures and their belief that 'information wants to be free' (Marwick, 2017). Referred to as 'free speech absolutism' (Nagle, 2017; Marantz, 2019), this form of extreme speech often features a fierce devotion to rational argumentation principles.[4] In this sense, it is not free speech that united the right on YouTube at this point in time, but rather a vernacular conception of free speech that found its apogee. Given its place as a signature issue for the New Left in the United States in the 1960s, this vernacular conception of free speech ironically led a number of YouTube's alternative influencers to position themselves as 'the NEW Counter-Culture' (Watson, 2017).

Lewis (2018) points to a particularly problematic trend within this new 'counter-culture', as far-right actors entered into debates with Alternative Influence Network (AIN) pundits around the topic of scientific racism, or 'race realism' as it is referred to. The concept of 'Race realism', as propagated today in its intellectualized form, links to older controversial work on race and intelligence—most famously Herrnstein and Murray's (1994) 'The Bell Curve'. The specific term was offered up in a 2005 paper by Canadian and American psychologists Rushton and Jensen, to describe the allegedly heritable IQ differences between races, which ideologically motivated social scientists were said to purposely neglect (Rushton and Jensen, 2005). Yet despite critical response to the scientific rigor of the concept, and widespread concern surrounding its politically motivated promotion, Lewis (2018) shows how, combined with YouTube marketing strategies such as guest appearances and the performativity of authenticity, the topic was able to attract new spectators under the header of free speech absolutism.

Still, Lewis' analysis remains mainly at the level of marketing tactics, or 'supply', and does not go so much into the way the discourse around scientific racism evolves as a dialogue between influencer and audience. Empirical contributions to Lewis' findings have suggested that it is not exclusively YouTube or even recommendations that are primarily responsible for the problem of extreme speech proliferating across the platform (Munger and Phillips, 2019; Ribeiro *et al.*, 2019). Based on quantitative analysis of audience engagement with AIN channels, Munger and Phillips (2019) claim that 'the novel and disturbing fact of people consuming white nationalist video media [i]s not caused by the supply of this media "radicalizing" an otherwise moderate audience' but rather, and more disturbingly, that it is a question of audience demand, previously constrained by the more limited scope of the ideology of extant mainstream

media (MSM) (p. 12). To consider audience demand, Ribeiro *et al.* and Munger and Phillips recommend paying attention to demands for extreme speech as a language that seeks to transgress and thereby 'widen' this previously restricted scope.

In the following sections, we present a method for capturing the supply and demand for extreme speech by using a combination of natural language processing techniques and digital methods (Rogers, 2013). This method consists in tracing the evolution of speech related to concepts or issues out of which norms arise—race, for example—and mapping the supply and demand for vernaculars intended to transgress them. We capture the exchange of scientific racism or 'race realism' vernaculars between content creators (channels) and their audiences (video comments) as one such kind of transgressive or 'extreme' discourse. While this method is reproducible on other datasets, our analysis speaks to a historical moment at which one of the largest Silicon Valley platforms doubled as an extreme speech marketplace.

3 The Dataset

Our dataset was initially collected by Dimitri Tokmetzis, data journalist and extremism expert at the Dutch news outlet *De Correspondent* (Tokmetzis, 2021). The dataset was collected throughout the fall of 2018, a period when the free market of extreme speech identified on YouTube was relatively ungoverned. Tokmetzis used the YouTube API v.3 to capture the YouTube channels of left- and right-wing political parties, media organizations, NGOs, and think tanks identified in Wikipedia, academic literature, right-wing extremist forums (4chan/pol, 8chan), and reports by anti-fascist NGOs, such as Hope not Hate and Kafka (Tokmetzis, 2019; Kafka, 2021). Ranging from 2006 to late 2018, his dataset contains 950 right-wing channels (see Supplementary Appendix II), with 253,621 video transcripts and 34,161,941 comments. The dataset also contains channels that the seed list has featured or subscribed to, including forty of the fifty-four channels identified by Lewis (2018).[5] As this dataset was larger than Munger and Phillips' (2020) and Ribeiro *et al.'s* (2019), it contains most of the significant right-wing channels in those datasets as well as smaller channels oriented in European politics.

Since this initial data collection, YouTube has become far more vigilant in policing 'harmful and supremacist content' via an expanded notion of 'hate speech' (YouTube, 2019)—which might be read as an acknowledgement of the 'textured nature of online abuse'. From March 2019, it began to sanction accounts for using 'racial, ethnic, religious and other slurs' or for 'making statements that one group is less than another based on [these] attributes, such as calling them less intelligent, less capable, or damaged' (ibid). In June of that year, it also released a blog post outlining its 'ongoing work to tackle hate' by targeting nonovert hate speech, including 'hateful and supremacist content' that denies 'that well-documented violent events like the Holocaust ... took place' or 'videos that promote or glorify Nazi ideology' (ibid). As a result, in the course of 2019 alone, 35% of the channels in our dataset had been taken down or 'deplatformed' (Rogers, 2020), which is to say that they no longer show up on the platform (Fig. 1). In the wake of this widespread 'deplatforming', far-right actors appear to have decamped across a diverse range of private servers and platforms branding themselves as the 'true' home of free speech like Parler and Gab, maintaining the affordances of digital networks perhaps at the expense of a mainstream vehicle for their 'alternative influence'.

4 The Method

Since the late 1990s, computer scientists have had to respond to a growing preoccupation with protecting vulnerable users or minorities from hateful language and 'cyberbullying' in messaging boards and late social media platforms (Davidson *et al.*, 2017; Gambäck and Sikdar, 2017; MacAvaney, *et al.*, 2019). In this respect, hate speech studies have focused primarily on 'detection', seeking to hone computational techniques more and more sensible to the social, linguistic, and contingent aspects of hate speech (Iyyer *et al.*, 2014; Kenter *et al.*, 2015; Azarbonyad *et al.*, 2017). Other studies have focused on the diversity of

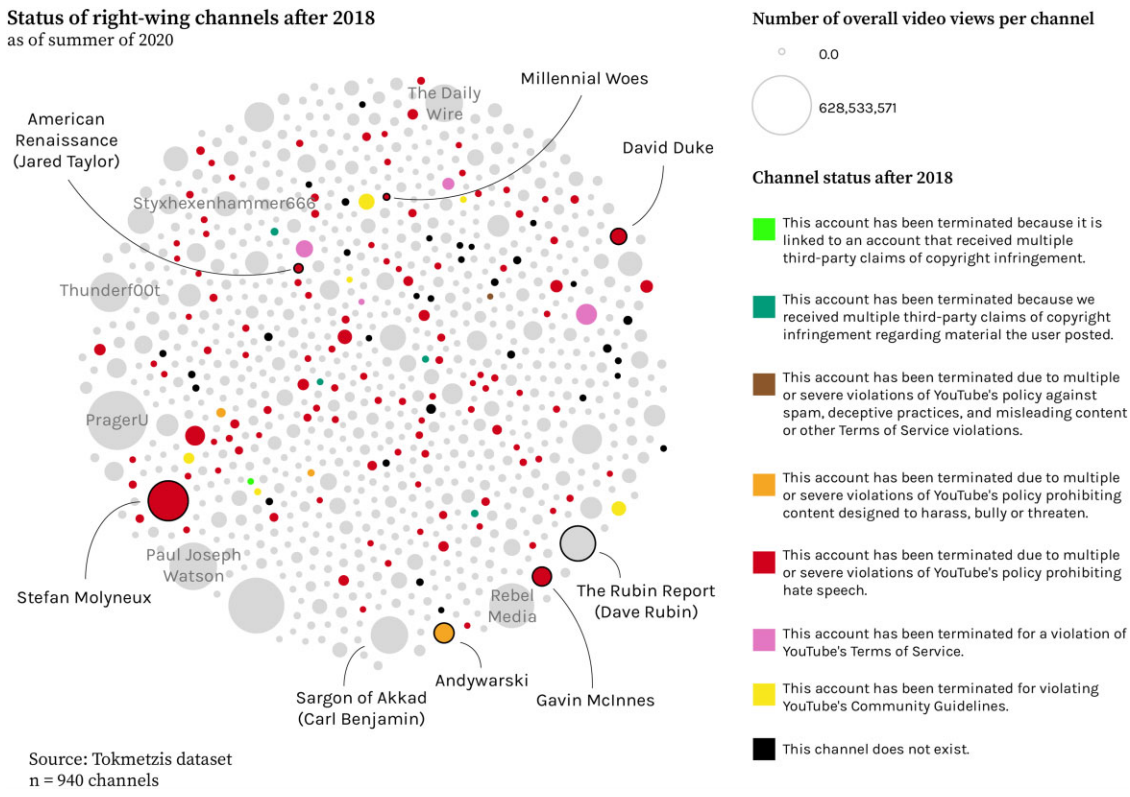

Fig. 1 Status of right-wing channels after 2018. Statuses were scraped from each of their YouTube pages in the summer of 2020. Guilherme Appolinário and Eleonora Cappuccio have contributed to previous versions of this image

definitions of hate speech, as well as the difficulty of subscribing to one over the other (Fish, 2020). Accordingly, a problem these studies raise is the continuous attuning of hate speech detection techniques to various 'subtleties in language' and 'interpretability problem(s)' (MacAvaney *et al.*, 2019).

As identified by Ribeiro *et al.* (2019), one yet under-researched avenue concerns tracing the discursive 'evolution' of extreme content, in which much of what is said in such discussions defies an 'easy binary division into speech that is acceptable and speech that is not' (Pohjonen and Udupa, 2017, p. 1174).[6] Indeed, many of the pundits studied by Lewis, Ribeiro *et al.*, and Munger and Phillips pepper their discussions with aspects of subcultural web vernacular—referring, for example, to their opponents as 'social justice warriors' (SJWs) or the 'MSM'. In YouTube debates on race-related issues, many of the vernacular terms used in the comment section are explicit 'alt-right' racist slurs, which are popular on the notoriously racist imageboard 4chan/pol (see Tuters and Hagen, 2020). Yet beyond these alt-right vernacular slurs, there also exists an 'intellectualized' language of 'highbrow white-nationalism' (Hawley 2017), echoing older eugenic discourse and other intellectual canons of far-right political philosophy. This type of scientific racist language poses a problem for content moderation, as it avoids the hostile vernacular forms usually found in hate speech despite still being discriminatory in substance.

In order to capture a wider spectrum of racist language, we sought to trace the overall evolution

of discussions related to 'race' in right-wing videos and comments (Fig. 2). Besides applying a list of hateful terms and operationalizing predictive techniques to capture subsequent hateful language varieties, we captured word collocations (bigrams) for the term 'race' in right-wing comments and video transcripts as a distant reading method for tracing the evolution of speech related to race (Figs 3 and 4). We used a window size of twenty words and obtained the thirty most frequent collocations per month. Using Rieder's 'Rank Flow' tool to 'distantly read' our data (Webber *et al.*, 2010; Rieder, 2021), we were able to identify a change—or 'evolution' in the framework of Ribeiro *et al.* (2019)—in how the concept of 'race' was framed in discussions over the course of 2015–18, from race 'baiting' to race 'realism'. Doing so resulted in filtering Dimitri's original dataset into a total of 253,621 videos and 34,161,941 comments.

Our next step was to engage in a close reading of both transcripts and comments in order to understand this pattern, as well as to try and determine what different uses of the word 'race' meant in different contexts. Results contained both (1) normative discourses around the meaning and right usage of this concept and (2) the types of discourses that offend said norms. The former includes debates about the uses and misuses of the word race, such as 'race baiting' or 'playing the race card', while the latter refers to intellectualized or scientific racist terms as well as hate speech slurs. At the risk of slightly complicating the analysis, the two aforementioned ways of talking about race led us, in fact, to produce three categories. We refer to the first of these categories as 'accusatory', the second as 'scientific racist', and the third as 'hate speech'. To obtain a picture of the absolute and relative prominence of each of these three discourses of race talk within our right-wing YouTube dataset, we then produced three representative word lists for each discourse and used them to automatically tag transcripts and comments (Figs 6, 8, and 10).

Word lists combined resulting bigrams with additional words from representative corpora (see Supplementary Appendix I). Scientific racist words included bigrams such as 'race realism' as well as forty-five keywords from Metapedia's Race and Intelligence article for 'Race and Intelligence' (de Keulenaar, 2019; Metapedia, 2021), a far-right wiki that is mostly constituted of passages from Encyclopedia Britannica's controversial eleventh edition of 1910–11 (Chalmers, 1992). Both encyclopedias use lexicons from late 19th-century racial theory, such as 'caucasoid', 'subspecies', 'race and IQ', or 'race taxonomy'. For hate speech, we used 103 words classified as 'racism' in the online hate speech database Hatebase.org and in Peeters *et al.*' (2020) list of hate speech vernaculars. While the former list includes words derived from public conversations classified with probabilistic linguistic analysis of hateful contexts (Hatebase, 2019), the latter contains words found in forums such as 4chan/pol, known for their far-right political culture and influence over online subcultural vernaculars (Peeters *et al.*, 2020). For posts mentioning 'race baiting', we used ten words that refer to race baiting or [playing the] race card, found only in our results.

Absolute and relative frequencies were not calculated in the same way for all word lists. As hate speech slurs may be uttered independently of the word 'race', we decided to calculate their absolute and relative frequencies as unigrams (that is, single words). Accusatory and scientific racist words must be mentioned close to the term 'race' to yield significant results; however, one may mention 'race' and 'iq' within close proximity without necessarily saying 'race and iq' as such. We have thus counted their absolute and relative frequency as bigrams, that is, words that must be within close proximity of the word 'race' (specifically, a window of twenty words). Frequencies were calculated for transcripts and comments separately. Due to their dramatic difference in size, they are displayed in different scales.

5 Limitations

Before proceeding, we should acknowledge a few methodological limitations. These are: coding our data; complementing gaps in our dataset; and data ethics.

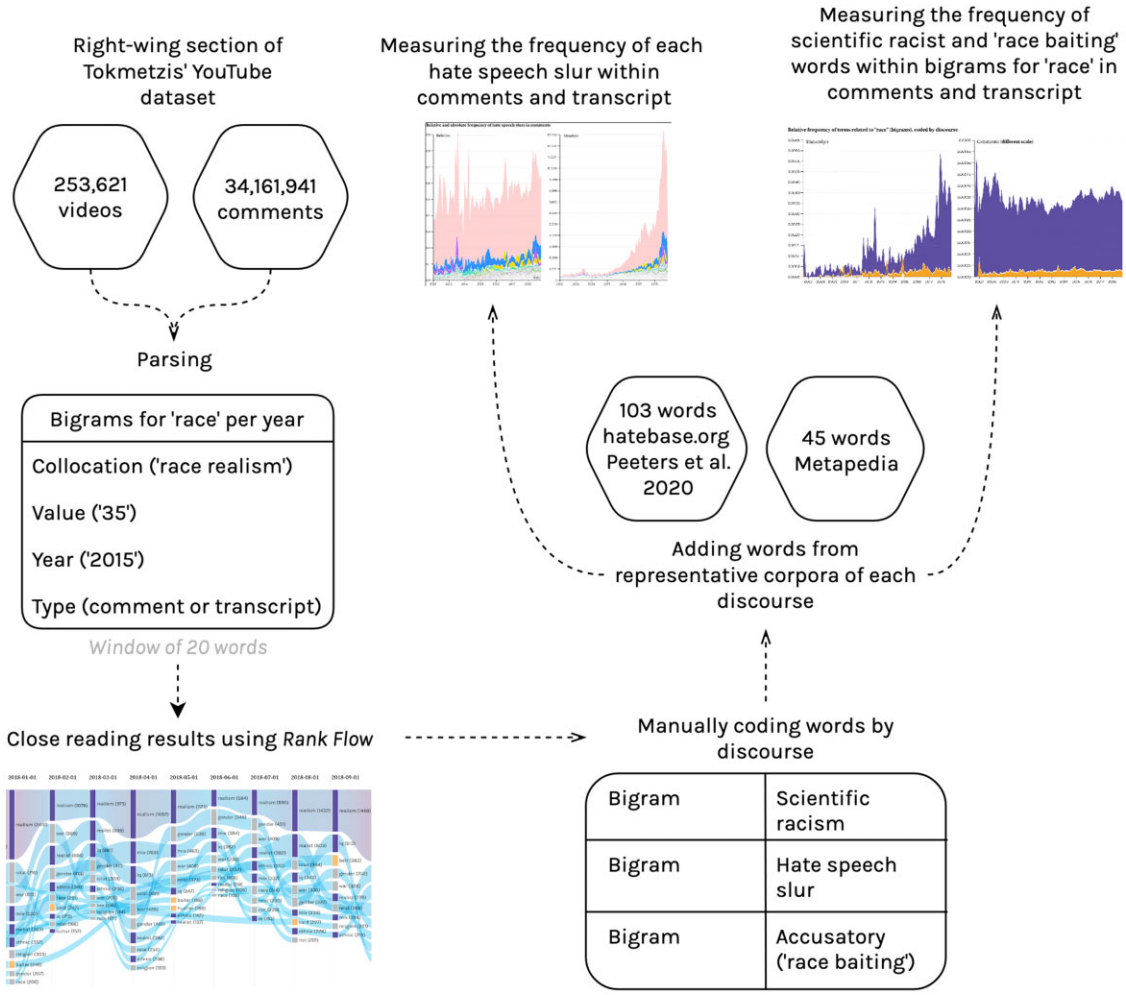

Fig. 2 Method diagram

In our analysis, we have sought to overcome the first challenge by using natural language processing techniques in combination with the qualitative methods informed by domain expertise. With mixed methods, we are able to highlight shifts in the discursive practices of YouTubers and their audiences with a distant reading of words related to the term 'race' in transcripts and comments. Distant readings are, however, unable to make claims about the intentions behind language use, even though the trends observed below do comport with those noted in YouTube radicalization literature. The coding we have used for additional distant reading, particularly our historical sample of scientific racist corpus, does not account for new terms that are arguably part of scientific racist discourse today; as a result, collocations like 'race war' are not part of this sample.

Also, our data have at times been incomplete by default: automatically generated transcripts on YouTube are not activated in all videos, nor do they capture unusual vernaculars often applicable to Internet hate speech linguo. Still, our dataset contained a number of right-wing channels, videos, and comments equivalent if in some cases larger than those of Ribeiro *et al.* (2019), Lewis (2018), and Munger and Phillips (2019). We were also able to

process words from Peeters *et al.* (2020) list in more than 20 million comments.

Finally, with the rising trend of computational analysis within the digital humanities, scholars do need to question the extent to which user data can be harvested and processed and for what ends. In addition to justifying our study in terms of the societal benefit gained from understanding extreme speech, our analysis also uses a distant reading approach to help ensure the privacy of the individual commenter. Where the analysis does refer to specific comments, for example, in a screenshot under a particular video, the commenters are anonymized.

6 The Analysis

Our analysis of YouTube's free market in extreme speech starts by examining the 'supply-side' of extreme speech around the topic of race, identifying two dominant discourses found in video transcripts. We then map the 'demand' for extreme speech by comparing the relative and absolute prominence of extreme speech terms in comments with those on transcripts (Figs 5 and 7). We conclude with a complementary analysis of hate speech slurs in comments to connect extreme speech with general levels of toxicity.

7 Supplying Extreme Speech

Close reading word collocations (bigrams) for the term 'race' in transcripts, from 2007 till 2018, revealed two types of discourses around race: (1) accusations of 'race baiting' or ['playing the] race card' and (2) 'scientific racism', evidenced by collocations such as 'race mix', 'race traitor', and 'race realist'. The interchange between these two discourses in 2015 paints a picture of general debates about the usage and meaning of the term 'race' (Fig. 3).

In the context of discussions about race, terms like 'race baiting' or '[playing the] race card' are often used to denote an alleged tendency by left-wing 'Social Justice Warriors' ('SJWs') as they were caricatured at the time, to introduce this topic in a debate as non-sequiturs. Close reading transcripts, we find that YouTubers refer to 'race baiting' or [playing the] race card' as attempts to cloud logic and facts by appealing to emotion through false or unwarranted accusations of racial discrimination. Such critiques feed into the idea that progressives rely on excessive emotional persuasion to discuss racial discrimination,

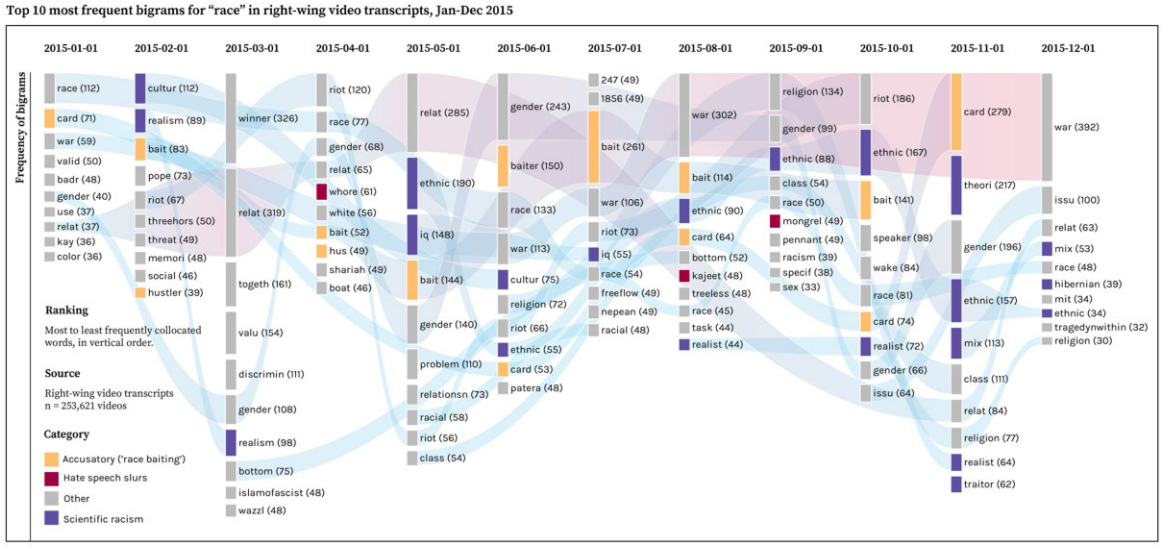

Fig. 3 Top ten terms most frequently associated with 'race' in 2015 right-wing transcripts (*X*-axis). Results are ranked most to less frequent from top to bottom and sorted per date (*Y*-axis)

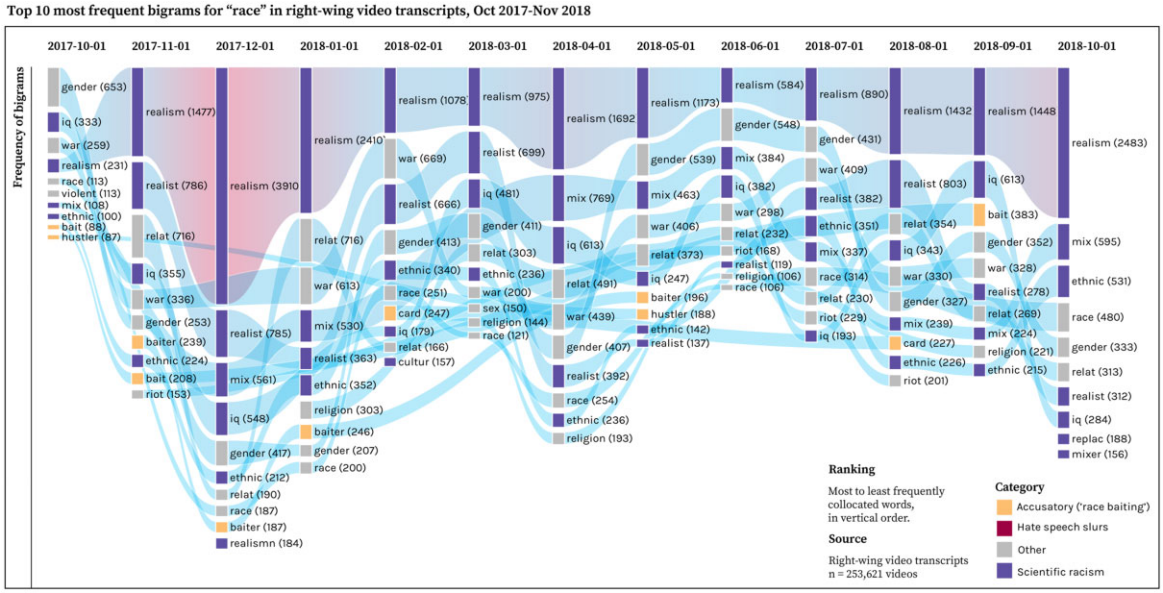

Fig. 4 Terms frequently associated with 'race' in right-wing video transcripts from October 2017 to November 2018 (*X*-axis). Results are ranked most to less frequent from top to bottom and sorted per date (*Y*-axis)

and thus lose touch with logical and factual reasoning. It should be noted that the usage of the term 'race baiting' in this way is a subversion of its dominant meaning, which according to the American Merriam-Webster dictionary (n.d.) is 'the making of verbal attacks against members of a racial group'.

The discourse around 'race' with accusatory terms such as 'race baiting' and 'race card' can be tied to two main arguments voiced by right-wing YouTubers at this time. First, it refers to the sentiment that progressives cannot be trusted to speak truthfully about issues around race, as their emotions trump logic and render impossible a consensus about what constitutes racism. Conversely, it frames 'race' as an issue that must take as its starting point a rationalist, indifferent approach. It is precisely this rationalist, allegedly 'scientific' approach that can be found by early 2017, when different collocations come to dominate right-wing discussions around the topic of race: 'realism' (Fig. 3). We thus present the observed difference between Figs 3 and 4 as evidence of 'the evolution of the speech of content creators' (Ribeiro *et al.*, 2019, p. 10) on right-wing YouTube channels, in which a formerly dominant accusatory frame of 'race baiting' is succeeded by the intellectualized framing of 'race realism'.

Figure 3 reveals that terms like '*race realis**' (race 'realist' or 'realism') were the most frequently associated terms to 'race' by the end of 2017.[7] As noted, 'race realism' references a highly controversial attempt among a small group of largely North American authors and political activists to reinstitute the academic study of race as a material (and thus biological) reality, rather than 'social construct', in a way that highlights allegedly 'natural' or genetic causes for racial inequality. By early 2018, race 'realist' discourse was the predominant way on YouTube to discuss issues around race and began to be associated with 'IQ', 'mix', 'riot', 'replacement', and later 'traitor'.[8]

Close reading an exemplary case in the right-wing section of our dataset, we find a 2017 interview with Stefan Molyneux,[9] in which he is asked by Dave Rubin from The Rubin Report about his controversial stance on the topic of race and IQ (Fig. 5). Molyneux starts by stating that he is not a scientist, but that based on scientific research he posts on his website, he finds '[race realism] unbelievably heartbreaking [and] one of the most difficult facts I ever had to absorb in my life'. Rubin interjects, stating: 'It is interesting that you are describing it as "heartbreaking" and "struggle", [and] hearing you frame it [as such] is actually

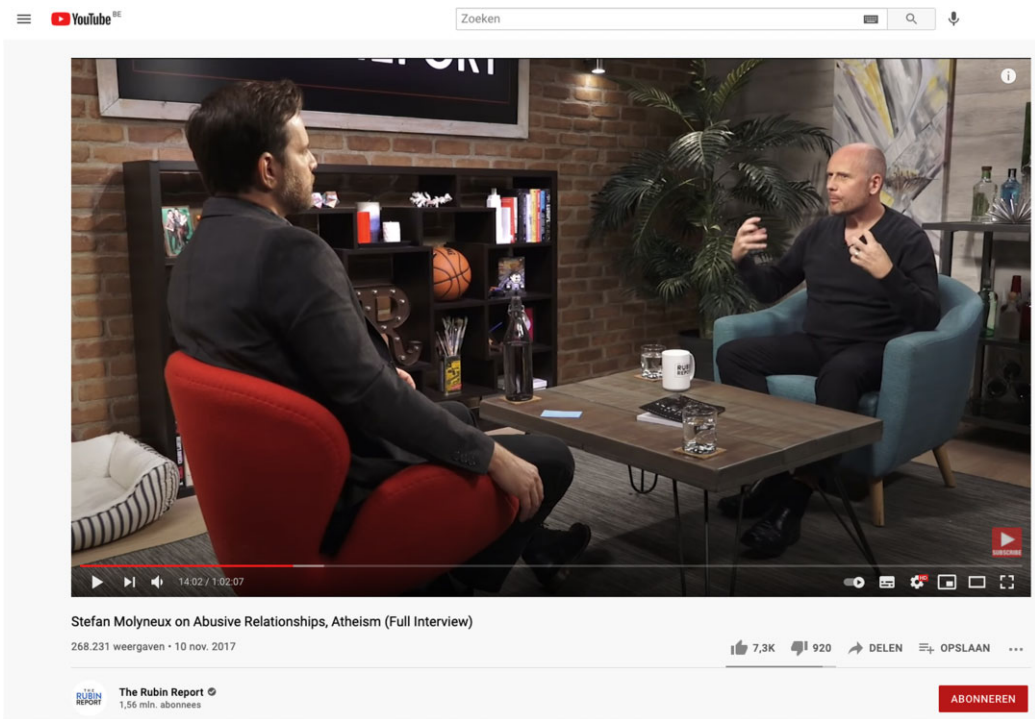

Fig. 5 Screenshot of The Rubin Report with Stefan Molyneux as guest (this screenshot was taken on 14-05-2021 and is, despite the banning of Molyneux's channel, still available to users on the platform)

different than the impression I had of you' (The Rubin Report, 2017). Molyneux's rhetorical style separates the 'scientific', 'rational', and 'logical' from the 'struggle' and 'heartbreaks', fitting the predominant evolution of 'race' discourse on YouTube around that time. Thus, Molyneux appears to have adapted scientific racist language to the 'race-baiting' critique that had been espoused in other videos of the time.

To get a sense of the magnitude of how scientific racism grew on YouTube, we measured the absolute and relative frequency of scientific racist language (as uni- and bigrams) and accusatory language ('race baiting') in transcripts between 2007 and 2018 (Fig. 6). We find that the period of interchange between accusatory ('race baiting') and scientific racist discourses found in Fig. 3 was in fact generalized across 2008 and 2015, before scientific racist terms significantly outgrew the accusatory language. This again suggests that an increase in extreme speech in our right-wing section arose in the context of contentious debates around how to carry on (bipartisan) conversations about race, particularly as these conversations polarize around conceptions of race as either a social construct or biological reality (Chou, 2017).

Figure 6 also shows that, beyond the absolute growth of far-right channels around 2017 (see Munger and Phillips, 2020) and the absolute growth of scientific racist terms, the relative frequency of scientific racist discourse picks up significantly by 2017.[10] This corresponds with a larger political culture of debating dangerous or taboo intellectual canons on race, as for example, in the 'Bloodsports' genre of discussion videos that Lewis suggests was key in introducing audiences to more extreme ideas (2018, p. 33). While the talking head, monological punditry adopted by a large portion of the channels continues over time, throughout 2017 and 2018 these 'discussion' videos trended on the platform. One of the debates perceived as most extreme at the time was the event noted at the start of the article, featuring figures such as Richard Spencer, Carl Benjamin, and Tarl Warwick (Styxhexenhammer666). Officially

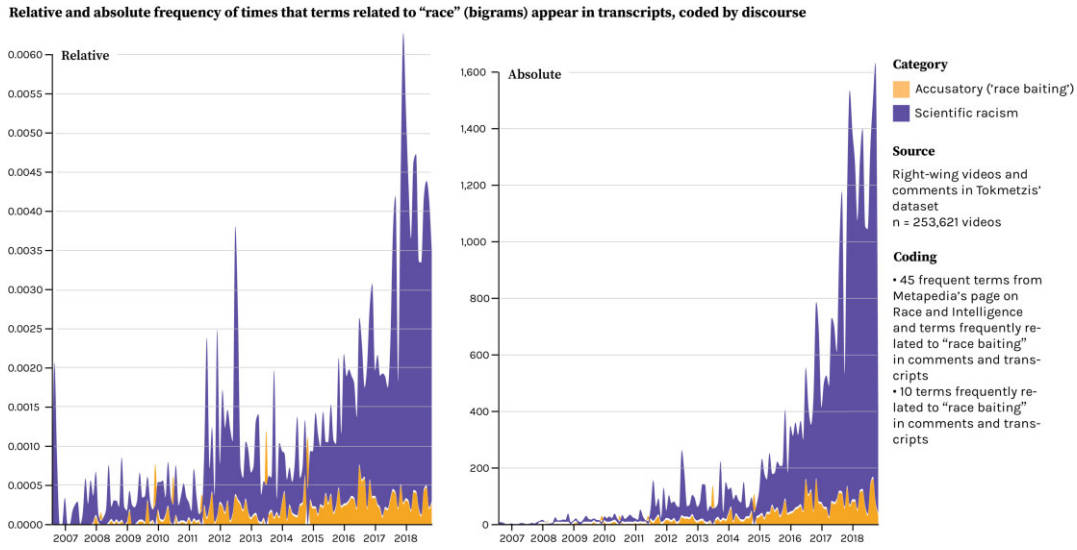

Fig. 6 Absolute and relative frequency of bigrams for the term 'race' in the transcripts of 253,621 videos from 950 channels, coded by discourse. Relative figures account for the exponential increase of videos in our dataset

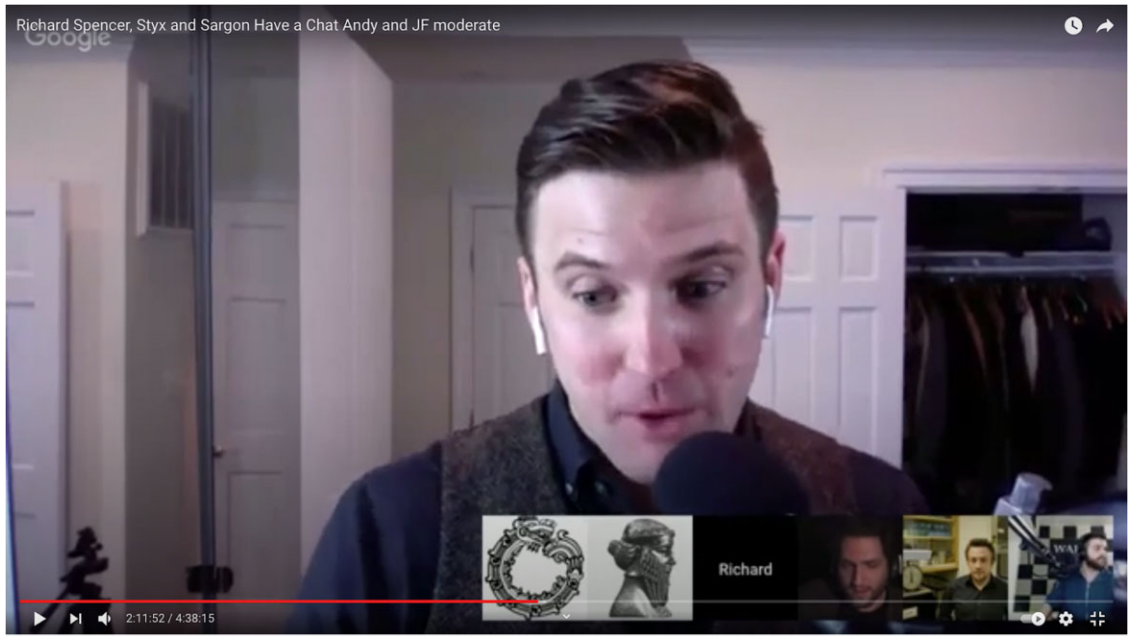

Fig. 7 Screenshot of Richard Spencer, Styx and Sargon Have a Chat (reuploaded by YouTube Politics Archive Channel on YouTube and thus still available to users on the platform)

titled *Richard Spencer, Styx and Sargon Have a Chat Andy and JF Moderate* (Fig. 7), the video had received over 600,000 views, 19,000 comments, 18,000 likes, and 1,800 dislikes at the time of our analysis, and despite having a runtime of over 4 h, was framed by channel host Warski as 'rebellious entertainment'.[11]

In her analysis of YouTube's AIN, Lewis claimed that more extreme (far-right) figures build their (online) reputation through guest appearances on YouTube shows of more popular and politically moderate figures. For instance, in relation to the video mentioned above, she portrayed Warski and Benjamin as major intermediaries, with Richard Spencer and fellow 'debater' Tarl Warwick as more peripheral nodes. While each of these figures had a channel with a relatively significant viewership (Fig. 1), Benjamin's was the largest at 1 and 400 millions, while Spencer's was by far the smallest at 24,000—Warski and Warwick had 240,000 and 390,000, respectively. Importantly, given the fetishization of intellectual performance that marked both the political microcelebrity genre more broadly and race realism as a topic, in this particular debate Benjamin was perceived to have lost. This had the effect of diminishing Benjamin's celebrity while at the same time increasing that of Spencer, with Spencer deeming the debate as the 'Unite the Right of YouTube' in reference to the Charlottesville march some months earlier (Lewis, 2018, p. 43).

8 Demanding Extreme Speech

Having established the shift in the supply of extreme speech around race and the relative increase in frequency, we mapped the demand side by applying our method to the comment sections of the videos in our dataset. Figure 8 shows the absolute and relative frequency of comments around 'race baiting' and 'scientific racism'.

Figure 8 shows comments that use any of the scientific racist uni- or bigrams from our list. While the absolute frequency of this language spiked as YouTube grew exponentially, an interesting finding here is that, in contrast to transcripts (Fig. 6), the relative frequency of scientific racist language over time remained relatively steady. The gambit here is that, to a greater or lesser extent, each of these terms may be understood as likely to occur in the context of discussions of scientific racist ideas, and more generally in undercurrent discussions about the role of race as an allegedly biological reality. From this distant point of view, we cannot fully account for the context

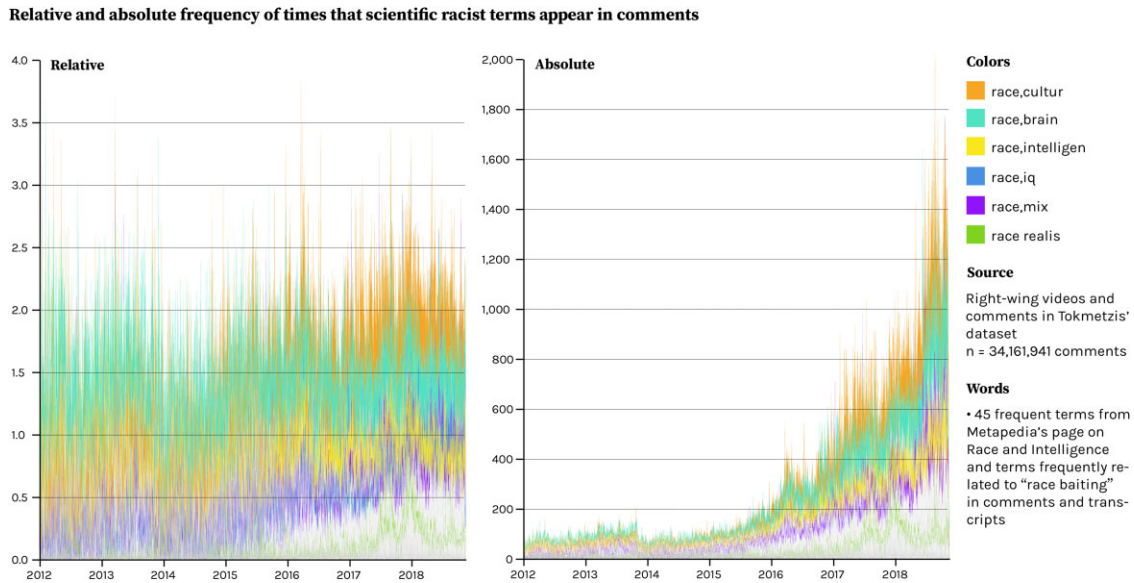

Fig. 8 Relative frequency of bigrams in 34,161,941 comments of 950 channels (overall), for the term 'race', coded by discourse. Words separated by commas are bigrams

of how these words were being used. Even though these are comments on right-wing channels, consistent with reception theory, there is likely a variety in the audience which would logically extend to the way that they use the terms. These terms could, for example, be used critically or humorously—though in many cases that too might be considered somewhat dubious. With that caveat in mind, in relation to Fig. 3, we offer Fig. 8 as possible evidence in support of Munger and Phillips' (2020) hypothesis. Earlier we framed the difference between Figs 3 and 4 as evidence of a shift—or 'evolution' in the phraseology of Ribeiro *et al.* (2019)—in how race is discussed in the right-wing video transcripts. Recalling Munger and Phillips' (2020) supply and demand hypothesis of YouTube radicalization, it is intriguing to interpret Fig. 8 as potential evidence that in discussing race realism, alternative influencers like Stefan Molyneux and Richard Spencer may in fact have been corresponding with a small yet consistent pre-existing audience demand.

Our analysis of extreme speech in the comments of right-wing videos can also be considered in relation to Munger and Phillips' (2020) observation that while there had been an overall decline in viewership of far-right YouTube videos since 2017, this did not affect audience engagement with those same videos (i.e. comment to view ratio)—which actually increased in that same period. How does the pattern of intellectualized extreme speech in Fig. 8 compare with the use of more overt hateful slurs by audiences of far-right YouTube videos? Utilizing the list of terms drawn from Hatebase.org and Peeters *et al.* (2020), in Fig. 9 we see a similar pattern to Fig. 8. Frequency of these terms in the comment section appears relatively steady over time, with a slight increase over time in terms that might also be associated with scientific racism (in blue), as well as in more explicit and transparent hate speech (in purple and yellow). Notably, one of these terms (in yellow) is also a recently coined slur that became popular on alt-right websites around this same period (Hawley, 2017).

In comparing Figs 8 and 9, we may conclude that the process of 'intellectualization'—as represented by the rise of race realism in the transcripts—does not replace more recognizable forms of race-related hate speech. We may treat these graphs as a sort of barometer for the state of race-related hatred on the platform during the period under study, which coincided with the rise of the 'alt-right'. As such, Fig. 9 seems to contradict the alt-right's self-image as a 'highbrow' discourse, as well as show that 'intellectualized' expressions of racism still appear alongside more vulgar forms of hate speech. In practice, it seems to expose the white nationalist intellectualization of a figure like Richard Spencer as dog whistles for an audience that includes its fair share of unashamed vulgar racists. Taken together, these two graphs would seem to support Lewis' more recent claim that demand for extreme political culture on YouTube is pre-existent, if merely dormant: the platform 'could remove its recommendation algorithm entirely tomorrow and it would still be one of the largest sources of far-right propaganda and radicalization online' (Lewis, 2020).

Finally, the question becomes how to interpret this pre-existent political culture on YouTube. Supplementing our quantitative work on the free market for hate speech and scientific racism, we performed a close reading of the comment section under the video 'Richard Spencer, Styx and Sargon Have a Chat' drawing on Lewis' (2018) discussion of the YouTube Bloodsports genre. Our analysis revealed that seemingly substantive 'debates' of ideas should be understood in relation to the exigencies of microcelebrity on the platform more broadly, and in the AIN in particular. This last point can be seen particularly well in the screenshot of the comment sections as shown in Fig. 10. The users who watched the debate and shared their interpretation did not necessarily engage directly with the various conceptions of a white ethnostate, as presented and defended by Spencer, but predominantly reflected on how well the microcelebrities had performed intellectually, particularly those that they were (previously) fans of. The top comment in Fig. 10, for example, echoes Spencer's ironic comments about Sargon's intelligence, or lack thereof, while others express their disappointment in his intellect. The second most popular comment at the time also expresses undercurrent affirmations about the biological reality of race, which Benjamin pays a high price for denying.

These moments, when a microcelebrity like Sargon is deemed intellectually inadequate by the YouTube audience to attack the ideas presented by Spencer,

appear important in the continuing demand for discussions about race, IQ and biology that figures such as Spencer are 'silenced' for. At the time of our analysis, this Bloodsport genre was part of a broader reactionary counter-culture in which such 'controversial' conversations offered radical microcelebrities like Stefan Molyneux a chance to insert themselves into the discussion and, consequently, build a following. While Lewis (2018) rightly points to the role of alternative influencers in the radicalization process of users as controversial figures are hosted by more established figures and introduced to a new audience, our close reading shows that the ability of these radical figures to outperform the other influencers 'intellectually', using the speech norms of the radical audience, is an essential part in the 'success' of the hosting session.

All in all, our close reading suggests a relationship of supply and demand within the microcelebrity genre that casts a new light on—or perhaps more accurately a shadow over—the much celebrated role of active audiences within participatory culture (Jenkins, 2008), as well as reiterating the notorious toxicity of YouTube's comment section and the appeal of online drama (Pihlaja, 2014). While we did not systematically drill down further to specific channels and videos in order to contrast various interpretive frameworks, we nevertheless feel confident that our findings may be interpreted as highlighting a pre-existing market demand, in line with Munger and Phillips' alternative thesis for YouTube radicalization, that shows the importance of the audience in the evolution of extreme speech and radical subcultures on YouTube.

9 Conclusion

This article has sought to contribute to scholarship on YouTube radicalization, specifically by taking up the proposal that future scholarship might 'trace the evolution of the speech of content creators and commenting users throughout the years' (Ribeiro *et al.*, 2019, p. 10). We did so by developing various approaches to study the 'textured nature of online abuse' (Pohjonen and Udupa, 2017, p. 1174), aspects of which may be of value for future moderation efforts within and across mainstream platforms. By focusing on the 'evolution'

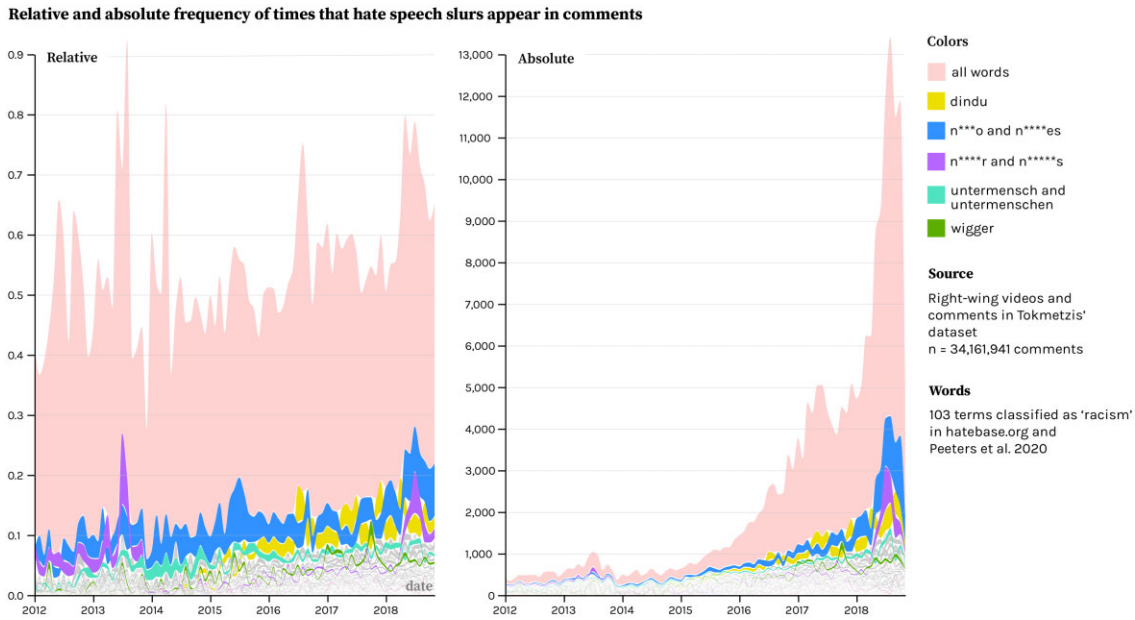

Fig. 9 Relative frequency of words from our list of hate speech slurs. A list of 103 words (see Supplementary Appendix I) were derived from 'racist' terms in the online Hatebase database (2019), and in Peeters *et al.* (2020) list of hate speech vernaculars

of 'extreme speech' our study moved beyond a methodological focus on recommender systems. The patterns that we revealed offer an empirical record of a significant moment in the platform's history. In light of YouTube's erasure of much of this data, these findings could not have been produced were it not for the foresight of Dimitri Tokmetzis, whose own analysis concluded that in the mid-to-late 2010's YouTube was 'the mothership of online hate, dwarfing obscure forums like 4chan and 8chan in size and influence' (Tokmetzis, 2019).

Our article has also built on Munger and Phillips' (2020) empirical study of YouTube radicalization. Based on a comparable though smaller dataset of right-wing YouTube channels, Munger and Phillips observed that, while extreme right viewership declined since mid-2017, engagement had increased. Based on bigrams and a word list of hate speech slurs we found a slight increase in our dataset (Fig. 10) that may substantiate that finding. Our more dramatic finding was the clear emergence of the concept of race realism in this same period. We found this pattern by looking at bigrams with race, in 253,621 video transcripts from 950 channels. In watching the videos and reading the transcripts in which race realism was discussed, we concluded that the concept amounted to a kind of 'intellectualization' of racial intolerance. In seeking to devise an approach to measure this phenomenon, we looked for terminology associated with racial science used in comments on race. Here, we found a steady and continuous pattern of co-occurrence over time. We considered this as possible evidence that 'intellectualization' of racial intolerance as represented in the emergence of race realism was already present in the comments. We thus concluded this interpretation could be read as evidence of Munger and Phillips' supply and demand hypothesis.

The theory that YouTube's right-wing 'alternative influencers' were supplying an underserved market demand for extreme speech should not be interpreted as letting YouTube off the hook as a mere intermediary, as platforms have so often argued in the past. As the infrastructure within which these exchanges take place, YouTube should also be considered culpable. If indeed YouTube can be considered to embody neoliberal ideals (Hokka, 2021), it should also be noted that neoliberalism views the free market as paradoxically artificial—a construction that must be consciously built from the ground up through a variety of material, technical and legal interventions (Foucault, 2008). While research is inconclusive as to the role of algorithmic recommendation in the radicalization process (cf. Ledwich), it nevertheless seems incontrovertible that in the late 2010s YouTube was indeed 'one of the most powerful radicalizing instruments of the 21st century' (Tufekci, 2018). Setting aside the affordance of algorithmic recommendation, we feel confident to conclude that far-right actors benefited from an environment structured by the affordances of the platform, notably an alternative network of microcelebrity pundits and opinion leaders, whose culture of 'debate' was crucial to the rise of the alt-right.

In retrospect, our analysis of race talk on YouTube represents a particular moment in YouTube's history when extreme speech proliferated across the platform. At this moment in time, an obsession with debate and logical pedantry that had long been present on the platform (Pihlaja, 2014) helped to turn it into an unregulated marketplace for extreme speech. In an effort to demonstrate their commitment to free speech, to prove their superior debating skills, and to win more audience, pundits who did not necessarily consider themselves as politically extreme nevertheless helped to create an environment in which such ideas as 'race realism' could flourish, and in which determined ideologues could masquerade as public intellectuals upholding free speech ideals. Indeed, several years later, a cohost of the Bloodsports debates, J. F. Gariepy, claimed to have used fellow host Warski as a 'puppet', and his channel specifically to legitimate race realism in front of millions of viewers (Gariepy, 2019). For Spencer's part, his seeming ascendency was curtailed shortly after the debate, with the release of a tape in which he could be heard screaming racist slurs in an emotional outburst following the 2017 Unite the Right March in Charlottesville. Fitting the dramatic internecine context of the competitive microcelebrity genre, the tape was leaked by Milo Yianopoulos, a former alt-right figure who had himself fallen from grace some years earlier. Amidst such infighting, the wider Bloodsports phenomenon would also soon dissolve, illuminating the dynamic nature of platform subcultures and the need for researchers to attend to shifts and currents rather than remain tied to more static forms of evaluation.

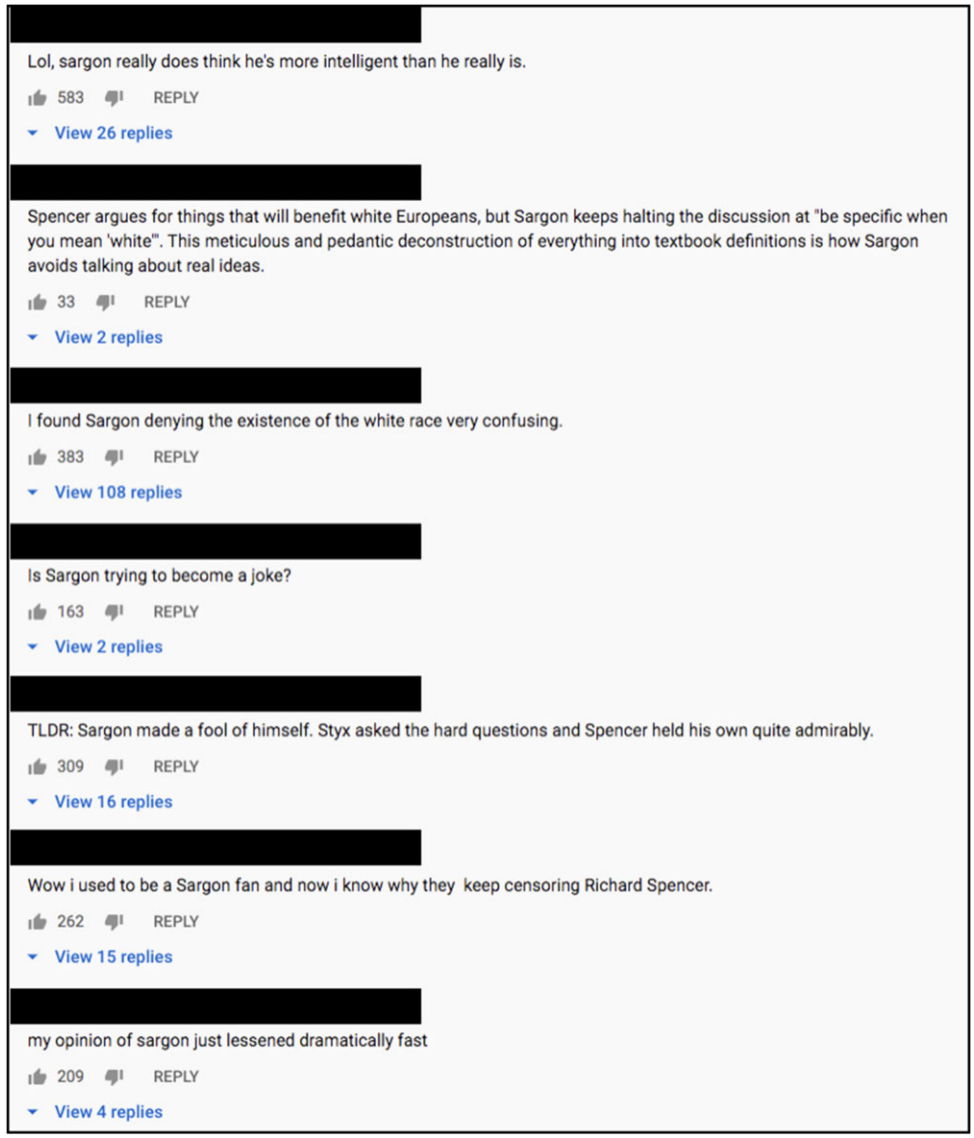

Fig. 10 Top comments from 'Richard Spencer, Styx and Sargon Have a Chat' (YouTube Politics Archive, 2020)

The problem of YouTube radicalization is too complex to be resolved through 'technological solutionism'. As we believe that our article has revealed, one important dimension in the broader problem of online radicalization involves the apparent resuscitation of discredited ideas which have historically been used to justify political oppression in the name of science. The kind of language emerging from these ideas, which we have referred to here as scientific racism, does not differ in substance from the types of hate speech slang platforms are usually trained to detect. Given the historical dynamics of discriminatory thinking and the likelihood that it continues to morph into different types of speech, we recommend that platform moderation focuses on targeting hateful contents by their substance rather than their form—that is, the ideas they express rather than

the ways in which they are expressed. Doing so would require monitoring the many ways in which discriminatory thinking evolves into various linguistic registers, be it in science, political philosophy, or colloquial language practiced by social media users.

Finally, it goes without saying that the liberal marketplace of ideas is not a level playing field. In addition to demographics and other structural factors, various platforms affordances attuned to user homogeneity and 'vanity metrics' (Rogers, 2013) also exacerbate structural inequalities. In practice, this means that discriminatory contents, unless moderated, may not encounter significant amounts of counter-ideas and demographics. Modifying existing information recommendation and filtering systems to facilitate the circulation of ideologically, linguistically, and demographically diverse information may help reduce the production of discriminatory contents as ideas and attitudes proper to fundamentally insular user cultures.

Supplementary data

Supplementary data are available at *DSH* online.

Funding

M. T. received financial support for this article from the ODYCCEUS Horizon 2020 Project (grant agreement number 732942) and from the AHRC funded Infodemic Project. E. dK. is supported by the UKRI-Canada ESRC Grant, *Responsible AI for Inclusive, Democratic Societies: a Cross-disciplinary approach to detecting and countering abusive language online* (ESRC reference ES/T012714/1). C. O.-C. received support for this article from the AHRC Grant-funded Project "Political Ideology, Rhetoric and Aesthetics in the Twenty-First Century: The Case of the 'Alt-Right'" (AHRC reference AH/R001197/1).

Conflict of interest

The authors declare no conflicts of interest.

References

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., and Kamps, J. (2017). Words are malleable: computing semantic shifts in political and media discourse. *arXiv:1711.05603 [cs]*. http://arxiv.org/abs/1711.05603 (accessed 17 February 2021).

Beran, D. (2019). *It Came from Something Awful: How a Toxic Troll Army Accidentally Memed Donald Trump into Office*. New York: All Points Books.

Bounegru, L., Gray, J., Venturini, T., and Mauri, M. (2017). *A Field Guide to Fake News: A Collection of Recipes for Those Who Love to Cook with Digital Methods (Chapters 1-3)*. SSRN Scholarly Paper ID 3024202. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3024202 (accessed 17 February 2021).

Chalmers, F. G. (1992). The origins of racism in the public school art curriculum. *Studies in Art Education*, **33**(3):134–43.

Chou, V. (2017). How science and genetics are reshaping the race debate of the 21st century. *Science in the News*, 18 April. https://sitn.hms.harvard.edu/flash/2017/science-genetics-reshaping-race-debate-21st-century/ (accessed 16 May 2021).

Davey, J. and Ebner, J. (2017). *The Fringe Insurgency: Connectivity, Convergence and Mainstreaming of the Extreme Right*. London: Institute for Strategic Dialogue.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv:1703.04009 [cs]*. http://arxiv.org/abs/1703.04009 (accessed 8 July 2021).

Dunphy, R. (2017). Can YouTube Survive the Adpocalypse?. *New York Magazine*, 28 December. https://nymag.com/intelligencer/2017/12/can-youtube-survive-the-adpocalypse.html (accessed 21 March 2021).

Feuer, W. (2019). Critics slam study claiming YouTube's algorithm doesn't lead to radicalization. *CNBC*. https://www.cnbc.com/2019/12/30/critics-slam-youtube-study-showing-no-ties-to-radicalization.html (accessed 27 May 2021).

Fish, S. (2020). *The First: How to Think about Hate Speech, Campus Speech, Religious Speech, Fake News, Post-Truth, and Donald Trump*. New York: Atria/One Signal Publishers. https://bookshop.org/books/the-first-how-to-think-about-hate-speech-campus-speech-religious-speech-fake-news-post-truth-and-donald-trump-9781508285175/9781982115258 (accessed 17 March 2021).

Foucault, M. (2008). *The Birth of Biopolitics: Lectures at the Collège De France, 1978-1979*. In Senellart Michel, (ed.),

Burchell, G. (tran.) New York: Palgrave Macmillan, pp. 267–289.

Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada: Association for Computational Linguistics, pp. 85–90.

Gariepy, J.-F. (2019). *Happy Birthday TPS! TPS 397*. https://www.bitchute.com/video/qsoYBi5a7oo/ (accessed 11 March 2021).

Google (2017) *Manage Super Chat and Super Stickers – YouTube Help*. Google. https://support.google.com/youtube/answer/7288782?hl=en-GB (accessed 11 March 2021).

Hatebase (2019). Hatebase, Hatebase. Available at: https://hatebase.org/ (accessed 17 March 2021).

Hawley, G. (2017). *Making Sense of the Alt-Right*. New York: Columbia University Press.

Hermansson, P., Lawrence, D., Mulhall, J., and Murdoch, S. (2020). *The International Alt-Right: Fascism for the 21st Century?* Abingdon, Oxon; New York, NY: Routledge (Routledge studies in fascism and the far right).

Hern, A. (2018). Facebook, Apple, YouTube and Spotify ban Infowars' Alex Jones. *The Guardian*, 6 August. https://web.archive.org/save/https://www.theguardian.com/technology/2018/aug/06/apple-removes-podcasts-infowars-alex-jones (accessed 17 February 2021).

Herrnstein, R.J. and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hokka, J. (2021). PewDiePie, racism and Youtube's neoliberalist interpretation of freedom of speech. *Convergence: The International Journal of Research into New Media Technologies*, **27**(1): pp. 142–60.

Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, MD: Association for Computational Linguistics, pp. 1113–22.

Jenkins, H. (2008). *Convergence Culture: Where Old and New Media Collide. Updated and with a New Afterword*. New York, NY: New York University Press.

Kafka (2021) *Antifascist Research*. Kafka. https://kafka.nl/ (accessed 14 March 2021).

Kenter, T., Wevers, M., Huijnen, P., and de Rijke, M. (2015). Ad Hoc Monitoring of Vocabulary Shifts over Time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, New York, NY, USA: Association for Computing Machinery (CIKM'15), pp. 1191–1200.

de Keulenaar, E., Tuters, M., Kisjes, I., and Beelen, K. (2019). On Altpedias: partisan epistemics in the encyclopaedias of alternative facts. *Artnodes*, **24**: 22–33.

Ledwich, M. and Zaitsev, A. (2019). Algorithmic extremism: examining YouTube's rabbit hole of radicalization. *arXiv:1912.11211 [cs]*. http://arxiv.org/abs/1912.11211 (accessed 17 January 2021).

Lewis, R. (2018). *Alternative Influence*. Data & Society Research Institute. https://datasociety.net/library/alternative-influence/ (accessed 9 December 2020).

Lewis, R. (2019). 'This Is What the News Won't Show You': YouTube creators and the reactionary politics of micro-celebrity. *Television & New Media*, **21**(2): 201–17.

Lewis, R. (2020). All of YouTube, Not Just the Algorithm, is a Far-Right Propaganda Machine. *Medium*, 8 January. https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430 (accessed 17 February 2021).

Livingstone, S. (2019). Audiences in an age of datafication: critical questions for media research. *Television & New Media*, **20**(2): 170–83.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: challenges and solutions. *PLoS One*, **14**(8): e0221152.

Marantz, A. (2019). *Antisocial: Online Extremists, Techno-Utopians, and the Hijacking of the American Conversation*. New York: VIKING, an imprint of Penguin Random House LLC.

Marwick, A. (2017). Are there limits to online free speech?. *Medium*, 5 January. https://points.datasociety.net/are-there-limits-to-online-free-speech-14dbb7069aec (accessed 17 February 2021).

Massanari, A. (2017). #Gamergate and the fappening: how Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, **19**(3): 329–46.

McCauley, C. and Moskalenko, S. (2008). Mechanisms of political radicalization: pathways toward terrorism. *Terrorism and Political Violence*, **20**(3): 415–433.

Merriam-Webster (2019). Race-Baiting. Merriam-Webster.com *Dictionary*. www.merriam-webster.com/dictionary/race-baiting (accessed 27 May 2021).

Metapedia (2021). Metapedia—countering semantic distortion worldwide. *Metapedia*. https://www.metapedia.org/ (accessed 8 July 2021).

Munger, K. and Phillips, J. (2019). A supply and demand framework for YouTube politics. *Preprint*. https://osf.io/73jys/download (accessed 28 October 2020).

Munger, K. and Phillips, J. (2020). Right-wing YouTube: a supply and demand perspective. *The International Journal of Press/Politics*, **21**(2). doi: 10.1177/1940161220964767 (accessed 28 October 2020).

Nagle, A. (2017). *Kill All Normies: The Online Culture Wars from Tumblr and 4chan to the Alt-Right and Trump*. Winchester, UK; Washington, USA: Zero Books.

Pariser, E. (2011). *The Filter Bubble: What the Internet is Hiding from You*. London: Viking.

Peeters, S., Hagen, S., and Das, P. (2020). Salvaging the Internet hate machine: using the discourse of extremist online subcultures to identify emergent extreme speech. *Zenodo*. doi: 10.5281/ZENODO.3676483 (accessed 20 February 2020).

Pihlaja, S. (2014). *Antagonism on Youtube: Metaphor in Online Discourse*. London: Bloomsbury.

Pohjonen, M. and Udupa, S. (2017). Extreme speech online: an anthropological critique of hate speech debates. *International Journal of Communication*, **11**: 1173–91.

Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A.F., and Meira, W. (2019). Auditing radicalization pathways on YouTube. *arXiv:1908.08313 [cs]*. http://arxiv.org/abs/1908.08313 (accessed 5 December 2020).

Rieder, B., Matamoros-Fernández, A., and Coromina, Ò. (2018). From ranking algorithms to 'ranking cultures': investigating the modulation of visibility in YouTube search results. *Convergence*, **24**: 50–68.

Rieder, B. (2021). RankFlow. *Digital Methods Initiative*. http://labs.polsys.net/tools/rankflow/ (accessed 28 March 2021).

Rogers, R. (2013). *Digital Methods*. Cambridge, MA: The MIT Press.

Rogers, R. (2020). Deplatforming: following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, **35**(3): 213–229.

Roose, K. (2019) The making of a YouTube radical. *The New York Times*, 8 June. https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html (accessed 27 May 2021).

Rushton, J. P. and Jensen, A. R. (2005). Wanted: more race realism, less moralistic fallacy. *Psychology, Public Policy, and Law*, **11**(2): 328–36.

Schmidt, E. and Cohen, J. (2014). *The New Digital Age: Transforming Nations, Businesses, and Our Lives*, **1**st edn. New York: Vintage Books, A Division of Random House LLC.

Steels, L. (2012). *Experiments in Cultural Language Evolution*. Amsterdam: John Benjamins Publishing Company.

Stokel-Walker, C. (2019). 'YouTube's plan to fix hate speech failed before it even started. *Wired UK*, 6 June. https://www.wired.co.uk/article/youtube-steven-crowder-ban-hate-speech (accessed 11 May 2021).

Sunstein, C. R. (2001). *Republic.com*. Princeton, NJ: Princeton University Press.

The Rubin Report (2017). Stefan Molyneux on Abusive Relationships, Atheism (Full Interview). Available at: https://www.youtube.com/watch?v=2-lN-KTpK_s (accessed 9 August 2021).

Tokmetzis, D. (2019). How they did it: exposing right-wing radicalization on YouTube. *Global Investigative Journalism Network*, 28 October. https://gijn.org/2019/10/28/how-they-did-it-exposing-right-wing-radicalization-on-youtube/ (accessed 17 February 2021).

Tokmetzis, D. (2021). decorrespondent/youtube_extremism. *De Correspondent*. https://github.com/decorrespondent/youtube_extremism (accessed 14 March 2021).

Tsukayama, H. and Timberg, C. (2017). 'Twitter purge' suspends account of far-right leader who was retweeted by Trump. *Washington Post*, 18 December. https://www.washingtonpost.com/news/the-switch/wp/2017/12/18/twitter-purge-suspends-account-of-far-right-leader-who-was-retweeted-by-trump/ (accessed 17 February 2021).

Tufekci, Z. (2018). Opinion | YouTube, the Great Radicalizer (Published 2018). *The New York Times*, 10 March. https://web.archive.org/web//https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html (accessed 5 December 2020).

Tuters, M. and Hagen, S. (2020) (((They))) rule: memetic antagonism and nebulous othering on 4chan. *New Media & Society*, **22**(12): 2218–37.

Watson, P. J. (2017). Conservatism is the NEW counter-culture [video file]. *YouTube*, 10 February. https://www.youtube.com/watch?v=avb8cwOgVQ8&feature=youtu.be&ab_channel=PaulJosephWatson.

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, **28**(4): 1–38.

Weisman, J. (2018). *((Semitism)): Being Jewish in America in the Age of Trump*, **1**st edn. New York: St. Martin's Press.

Weiss, B. (2018). Opinion | Meet the Renegades of the intellectual dark web. *The New York Times*, 8 May. https://web.archive.org/web/20190430030624/https://www.nytimes.com/2018/05/08/opinion/intellectual-dark-web.html (accessed 17 February 2021).

Wendling, M. (2018). *Alt-Right: from 4chan to the White House*. London: Pluto Press.

YouTube (2019). Our ongoing work to tackle hate. *Official YouTube Blog*. https://youtube.googleblog.com/2019/06/our-ongoing-work-to-tackle-hate.html (accessed 17 December 2019).

Youtube Politics Archive (2020). Richard Spencer, Styx and Sargon Have a Chat Andy and JF moderate. Available at: https://www.youtube.com/watch?v=Nnts-ezAE9o (accessed 9 August 2021).

Zuckerberg, D. (2018). *Not All Dead White Men: Classics and Misogyny in the Digital Age*. Cambridge, MA: Harvard University Press.

Zuckerberg, M. (2013). Is connectivity a human right?. *Facebook*. https://www.facebook.com/isconnectivityahumanright (accessed 2 March 2021).

Notes

1 The research discussed here originated from a project performed at the Winter School of the Digital Methods Initiative based at the University of Amsterdam. Its results are largely attributed to the research of project participants Laurie Le Bomin, Jonathan Hendrickx, Kristina Herbst, Mikkel J. Hjelt, Bart Josten, Magnus Knustad, Valérie van Mameren, Stephanie Tintel, and the authors Emillie de Keulenaar, Ivan Kisjes, Daniel Jurg, Cassian Osborne-Carey, and Marc Tuters.

2 Radicalization is defined here as an increased openness and commitment to out-group conflict and a reciprocal demand for in-group defense, which manifests as changes in belief (cf. McCauley and Moskalenko, 2008).

3 Munger and Phillips (2019) identify five ideological positions within the AIN: liberal (associated with the likes of Joe Rogan and 'Destiny'), skeptic (associated with the likes of Jordan Peterson and Carl Benjamin), conservative (associated with the likes of Steven Crowder and Ben Shapiro), alt-lite (associated with the likes of Paul Joseph Watson and Stefan Molyneux) and alt-right (associated with the likes of Richard Spencer and Red Ice TV).

4 Note that in the online forum 4chan/pol/—which has widely been identified as an essential meeting place for more extreme elements of alt-right political discussion (Nagle, 2017; Hawley, 2017; Wendling, 2018; Beran, 2019)—a PDF listing logical fallacies is permanently affixed as the first post that one sees when visiting the site.

5 Included in the dataset are such prominent figures in Lewis' AIN as: Faith Goldy, Lana Lokteff, Tara McCarthy, Tim Pool, Bre Faucheux, Mark Collett, James Allsup, Brittany Pettibone, Baked Alaska, Styxhexenhammer666 (Tarl Warwick), Andy Warski, Stefan Molyneux, Gavin McInnes, Mike Cernovich, Jordan Peterson, Roaming Millennial (AKA Lauren Chen), Millennial Woes (AKA Colin Roberson), Paul Joseph Watson, Steven Crowder, Lauren Southern, Tommy Robinson, Sargon of Akkad (AKA Carl Benjamin), Jared Taylor, Jean-François Gariépy, Caolan Robertson, and Black Pigeon Speaks (Felix Lace).

6 Here we also refer the reader to computational linguist Luc Steels' (2012) notion of 'cultural language evolution', which posits a coevolutionary relationship between cultural language and the 'pragmatics' of its context of use.

7 On 18 December 2017, the Twitter account of the 'highbrow white nationalist' and 'race realist' YouTube host Jared Taylor was suspended, as well as the account for his American Renaissance institute (Tsukayama and Timberg, 2017).

8 In line with the revelation that 'race' often co-occurs with 'realism', we also found that other terms most frequently collocated with 'race' were those from our sample of scientific racist language—as extracted from Metapedia's Race and Intelligence article (see Supplementary Appendix I)—in comparison to allegations of 'race baiting' as of 2017 (see Fig. 1).

9 Stefan Molyneux is an online influencer that has been labelled an 'alleged cult leader who amplifies "scientific racism", eugenics and white supremacism to a massive new audience' by the Southern Poverty Law Center. He was deplatformed in 2020 from YouTube due to violating their terms of service.

10 YouTube played host to several early online debates involving 'race realism' as suggested by spikes in 2012–

13, and indicating earlier forms of 'contrarian' interplay amidst skeptics, libertarians and political pundits. However, as illustrated, the issue failed to spread across the platform or gain significant attention until several years later.

11 It is important to note here that while much of the recent literature on the rise of the alt-right as a movement has focused on the role of anonymous online forums, such as 4chan and 8chan as the source of its style and creativity (Nagle, 2017; Beran, 2019), Lewis has more recently developed the counter argument that the movement has also been held together, to a large extent, by these YouTube microcelebrities (Lewis, 2019). Indeed, consistent with Lewis' observations concerning the role of YouTube microcelebrity in the alt-right, this particular debate was also much-discussed within the anonymous message board 4chan/pol/.