Digital Journalism

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rdij20

Accounting for Personalization in Personalization Algorithms: YouTube's Treatment of Conspiracy Content

Roan Schellingerhout, Davide Beraldo & Maarten Marx

To cite this article: Roan Schellingerhout, Davide Beraldo & Maarten Marx (2023): Accounting for Personalization in Personalization Algorithms: YouTube's Treatment of Conspiracy Content, Digital Journalism, DOI: 10.1080/21670811.2023.2209153

To link to this article: https://doi.org/10.1080/21670811.2023.2209153

Published online: 25 May 2023.

DIGITAL JOURNALISM
https://doi.org/10.1080/21670811.2023.2209153

Routledge
Taylor & Francis Group

ORIGINAL ARTICLE

OPEN ACCESS Check for updates

Accounting for Personalization in Personalization Algorithms: YouTube's Treatment of Conspiracy Content

Roan Schellingerhout[a], Davide Beraldo[b] and Maarten Marx[c]

[a]Department of Advanced Computing Sciences, Maastricht University, Maastricht, The Netherlands; [b]Department of Media Studies, Faculty of Humanities, University of Amsterdam, Amsterdam, The Netherlands; [c]IRlab, Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

This article investigates under which video watch conditions YouTube's recommender system tends to develop a preference for conspiracy-classified videos. Whereas existing research on so-called filter bubbles and rabbit holes tends to rely on non-personalized recommendations and on standard watch patterns, this study puts personalization and diversified user strategies at the center of its design. 20 authenticated bots have been instructed to watch YouTube content based on four distinct watch strategies. In a baseline strategy, bots watched non-conspiracy videos only. Treatment strategies involved watching conspiracy-classified content, selected based on either non-personalized, partly-personalized, or fully-personalized input. Bots watched a total of 15 videos, and after each video their top 20 homepage recommendations were collected and classified as either conspiracy-related or not. This allowed us to measure the impact of each video watched and of each watch strategy on the proportion of conspiracy-classified content recommended at each step. The same experiment has been reverted, exposing the treatment groups to non-conspiracy videos only, to assess the persistence of this pattern. Our results show that users primed with conspiracy-classified content tend to quickly receive a much larger proportion of conspiracy-classified recommendations. Inverting this pattern proves significantly more difficult than generating it. There are also indications that watch strategies relying on personalized content as input might produce stronger effects. This article contributes evidence to the argument that YouTube's recommendation system is prone to generating strong, potentially pernicious recommendation patterns. Moreover, it contributes a replicable methodology that puts personalization at the center of the stage in the study of content personalization algorithms.

KEYWORDS
Filter bubbles; recommender systems; personalization; YouTube; machine learning

CONTACT Roan Schellingerhout roan.schellingerhout@maastrichtuniversity.nl

Introduction

With an average of 34.6 billion page views per month, YouTube represents the world's largest video-sharing platform and the second most popular website on the internet (Neufeld 2021). Whether related to lifestyle, music, fiction, or politics, an estimated 70% of the videos watched are discovered by users through YouTube's recommender system (Cooper 2020). Considering the platform's centrality in contemporaries' information diets, it is important to put its recommendation system under scrutiny in the context of broader calls for algorithmic transparency and fairness (Gillespie 2014; Milan and Agosti 2019; Sandvig et al. 2014; Pasquale 2015). YouTube's recommendation system has already been at the center of attention and criticism because of its alleged tendency to spiral into bubbles or "rabbit holes" of toxic content (Tufekci 2018). Whereas YouTube's policy statement (YouTube 2021) promises to actively remove harmful content, and actions in this sense have been taken, the platform is still home to a community of channels promoting conspiracies, "alternative facts," and/or misinformation alike (Allington and Joshi 2020; Faddoul, Chaslot, and Farid 2020).

Arguments such as "filter bubbles" (Pariser 2011) and algorithmic "rabbit holes" (Tufekci 2018) have firmly entered the public debate. These arguments have for long relied on largely anecdotal accounts and might underestimate the relative importance of individual selective exposure and social processes of homophily (Bruns 2019; Hosseinmardi et al. 2020; Spohr 2017; Zuiderveen Borgesius et al. 2016). However, a growing body of empirical studies on the effects of recommender systems is emerging, including on YouTube (Airoldi, Beraldo, and Gandini 2016; Bryant 2020; Hosseinmardi et al. 2020; O'Callaghan et al. 2013; Ledwich and Zaitsev 2020; Kaiser and Rauchfleisch 2020; Romano et al. 2021; Roth, Mazières, and Menezes 2020), thus far providing mixed evidence. Most empirical research, moreover, focused on non-personalized, rather than personalized recommendations – an evident limitation for research on the effects of recommendation algorithms, which in real-world scenarios heavily rely on personalization.

Wishing to intervene in this pressing and debated issue, this article presents the findings of a study based on an original experimental setting, looking at recommendations directed at users engaging with conspiracy-classified content on YouTube. Following a sock-puppet approach to algorithmic auditing (Sandvig et al. 2014), brand-new Google accounts have been developed and programmatically directed to play YouTube videos according to different "watch strategies," simulating different ways users can engage with conspiracy-related content based on more or less personalized recommendations. After each video watched, the recommendations provided on the homepage of the user were scraped and labeled as being either conspiracy-related or not by a machine learning classifier trained on an existing, manually labeled dataset (Ledwich and Zaitsev 2020). Consequently, it was possible to assess the proportion of conspiracy-classified content at each step, evaluating the impact of different watch strategies against the baseline and its dynamics. The research design corresponds to an experimental setup where the proportion of recommended videos classified as conspiracy is the outcome variable, the number of videos watched at each measurement is the within factor, and the watch strategy followed by the bots is the between factor.

The article provides original empirical evidence related to the following question: *How does YouTube's recommendation system treat conspiracy-classified content based on personalized watch patterns?* In other words, the research assessed whether, and under which conditions, YouTube's recommender system tends to develop a preference for conspiracy videos over non-conspiracy ones, where "preferring" is defined as the situation in which the amount of conspiracy videos present in the recommendations is significantly higher than that of a baseline. Whereas we opted to look into conspiracy content because of its salience and controversial nature on YouTube, the main goal of the research is largely independent of this specific topic and speaks more in general to how the platform's recommender systems can be influenced by preferences for any type of specific content signaled by certain watch patterns. Different watch strategies have been implemented in order to account for different user behavior; the analysis also focused on the dynamics of the process, as important components of the problem are how fast a preference for certain types of content is developed, and to what extent such a preference is (if at all) forgotten. Hence, the study focuses on the following sub-questions:

- RQ1: Do watch strategies signaling interest towards conspiracy content increase the proportion of conspiracy content being recommended?
- RQ2: How Persistent Are These Effects?
- RQ3: How Do Specific users' Watch Strategies, Making Use of More or Less Personalized Inputs, Influence These Recommendation Patterns?

In broad terms, the study measured the tendency for YouTube recommendations to spiral down into a filter bubble of conspiracy content, as well as the persistence of such potentially pernicious patterns. The unit of analysis is actual recommendations collected over fifteen iterations, through twenty sock-puppet Google accounts simulating four different watch strategies. We are aware that demonstrating the existence of so-called "filter bubbles" or "rabbit holes" would require further conceptual elaboration as well as different types of data and analyses. Nonetheless, this article contributes original empirical evidence to the debate, supporting the claim that conspiracy content gets easily, quickly, and persistently preferred once a personalized watch pattern signals a preference in this sense.

The article is structured as follows. The literature section introduces the urgency of the issue, the ambiguity of existing results, and highlights the originality of this study. The methods section describes the data collection process, presents the setup of the experiments, and offers an overview of the analysis. The results section showcases our findings in relation to dynamics and relational patterns of recommendations. The discussion reflects upon the implication of the findings, their relation to previous findings, and the limitations of the present study. The conclusion summarizes the key points of the article.

To reproduce or extend our study, all described code and data are available on the public GitHub repository https://github.com/Roan-Schellingerhout/YouTube_conspiracy_paper.

Algorithmic Personalization on YouTube

The Public Relevance of Recommender Systems

Algorithmic systems mediate social processes unfolding online in a variety of ways (Gillespie 2014) and are at the center of contention in contemporary societies (Beraldo and Milan 2019). Research, advocacy, and activism around the issue of algorithmic transparency and accountability have thus spurred in the past years (Noble 2018; Pasquale 2015; Sandvig et al. 2014). One of the key issues associated with the algorithmic mediation of social life is that of algorithmic personalization (Milan and Agosti 2019), i.e., the way in which complex, often proprietary technologies such as filtering, ranking, and recommender systems influence users' exposure to information by tailoring their output to inferred (and, sometimes, engineered) individual preferences.

Sources and content that are algorithmically selected to maximize users' satisfaction (often measured as the time a user spends engaging with content) create personalized informational universes on spheres of life that range from leisure activities to political participation. The notorious "filter bubble" hypothesis (Pariser 2011) claims that this results in users being trapped in streams of content confirming their existing beliefs, with detrimental effects in terms of political polarization and the spread of misinformation. Whereas algorithmic-driven personalization can produce benefits in terms of efficient information retrieval and user satisfaction, the fragmented and self-reinforcing information universes they produce can also pose threats to social cohesion (Whittlestone et al. 2019).

The observation that algorithmic "pre-selected" personalization is generally paired with a conscious "self-selected" one, and the reliance of the filter bubble thesis on largely anecdotal accounts, have left some scholars wondering whether societal concerns around the issue have been exaggerated (Zuiderveen Borgesius et al. 2016). Selective exposure to information sources (Frey 1986), predisposition to confirmation bias (Nickerson 1998), and homophily in social networks (McPherson, Smith-Lovin, and Cook 2001) are all well-researched phenomena that pre-exist the advent of the internet. However, the fact that ideological polarization might depend to a larger extent on pre-existing individual or social patterns (Hosseinmardi et al. 2020) does not eliminate risks associated with algorithmic ones (Roth, Mazières, and Menezes 2020). Moreover, pre-selection and self-selection are interconnected with each other – more self-selection of specific content can trigger an increase in pre-selection of said content and vice versa. Therefore, either type of content selection should never be studied in a vacuum, as that will lead to an inadequate understanding of the algorithm. The inherent link between the two selection types makes it clear that previous works have been limited in their approach, exclusively viewing the type of selection as a binary (Roth, Mazières, and Menezes 2020; Hosseinmardi et al. 2020; O'Callaghan et al. 2013).

YouTube's Recommender System and Conspiracy Videos

Conspiracy content has been booming on YouTube (Donzelli et al. 2018), and the platform's recommender system is often understood as prone to the generation of "rabbit holes" – recommendation patterns that spiral down into increasingly extremist or generally problematic content. "Alternative news" and conspiracy channels have attracted growing audiences, allegedly contributing to political extremism and general distrust in mainstream media as well as in science (Tufekci 2018). Whenever this increased distrust relates to crucially important topics, such as believing in the efficacy of vaccines or in the legitimacy of an electoral process, it can create genuine dangers to the public (Rosenbaum 2021). An antidote to polarization and radicalization patterns could be that of diversifying the content presented to users, in order to challenge someone's viewpoints and avoid reinforcing their existing beliefs (Bozdag and van den Hoven 2015). However, since the platform's business model capitalizes on users' attention and engagement with content, YouTube's algorithm is broadly tuned to recommending videos that are likely to generate a lot of watch time (Maack 2019). As it turns out, controversial content (such as conspiracies of all kinds) tends to have higher audience retention: people keep watching controversial content for longer (Birch 2019). Whenever content is surprising (and conspiracy theories often are), it is more likely to capture and maintain a user's attention. Thus, by showing the user more diverse content, the system would actively hinder its own goals. Because of this design, "conspiracy filter bubbles" could be an endogenous outcome of a recommender system tuned to maximize users' engagement and adopt estimated watch time as the key factor for its selections (Chitra and Musco 2020). Especially individuals that are already more predispositioned to be interested in such content (e.g., because of their social environment, political affiliation, education) could be susceptible to the effects of such phenomena. For example, conspiracy theories have been shown to be more potent in convincing those subscribing to right-wing authoritarianism Frischlich et al. (2021), those with lower intelligence Furnham and Grover (2022), and those in a less-favorable socio-economic position Freeman and Bentall (2017).

Controversies around platforms' responsibilities towards content moderation and recommendation (Gorwa, Binns, and Katzenbach 2020) have led to the announcement of stricter policies. YouTube has ambiguous rules with regard to the spread of conspiracy videos on the platform (YouTube 2021). As long as the content does not directly incite violence or endanger public health (e.g., misinformation about the COVID-19 virus), misinformation is allowed to be shared. As a result, YouTube is home to multiple conspiracy communities, and videos related to theories such as "the earth is flat and the government is hiding it from us" or "the world is ruled by cannibalistic satanic pedophiles" gather millions of views on the platform (Paolillo 2018; Miller 2021). These communities not only harbor on YouTube itself, but also on so-called "dark websites," such as Gab and 8Kun (formerly 8chan) – platforms that receive limited moderation, causing them to be safe havens for extremist content (Zeng and Schäfer 2021). However, such websites often have limited resources, which leads to a lack of video hosting functionalities. As a result, YouTube is the most-cited website on dark websites, functioning as an intermediary platform. Considering most legacy media have stricter policies on conspiracy content in place, this makes YouTube a sort of safe haven for such content, which may contribute to the behavior of its algorithm.

YouTube's algorithmic recommendations have a remarkable influence on users' content consumption (Cooper 2020). YouTube's recommender system tries to suggest videos based on the expected watch time they will generate, rather than the probability of a user clicking on them (Covington, Adams, and Sargin 2016), a decision made in order to decrease the likelihood of deceptive, clickbait videos being recommended. In order to keep the user on the website as long as possible, which is profitable for YouTube, the algorithm prefers recommending videos that it suspects the user will watch for a longer period of time, allegedly even when they may contain harmful or otherwise toxic content (Maack 2019; Tufekci 2018). YouTube's recommender system makes its decisions by combining the similarity of content and signals from users' watching behavior. According to YouTube (Ledwich and Zaitsev 2020), a user's viewing behavior is responsible for approximately 70% of their recommendations.

Existing Studies and Contribution

Existing research on recommendation patterns on YouTube shows mixed results. This might stem from the great variability in research design, combining different sources of data (YouTube's API or users' interface), units of analysis (channels or videos), recommendations type (watch-next or homepage), recommendation depth (one or more levels) and modeled user (authenticated or not).

Roth, Mazières, and Menezes (2020) analyzed confinement patterns within networks of non-personalized, "watch-next" video recommendations linking channels. According to their findings, these recommendations quickly lead to a decrease in information diversity and tend to produce recommendation patterns more and more homogeneous in terms of broad topics (e.g., politics, music, entertainment). They also speculate that, as soon as the algorithm collects and includes information about a user in its recommendations, personalized recommendations could lead to an even stronger limitation of recommendations' variety. Similarly, Romano et al. (2021) found that watch-next recommendation patterns generated by (non-logged) users primed to signal interest towards either progressive or conservative channels, produce differences in terms of the type of sources and content recommended.

In the context of extremist bubbles, research based on large-scale longitudinal data of real users' browsing behaviors (Hosseinmardi et al. 2020) found evidence for a small but growing echo chamber of far-right content; however, a preliminary analysis on the causes attributes this outcome more to individual consumption patterns than to YouTube's recommendation system. A study relying on data obtained from YouTube's API, instead, provided evidence that users accessing extreme right videos are likely to be recommended further extreme right content (O'Callaghan et al. 2013). Similarly, in their large-scale audit of radicalization patterns on YouTube, Ribeiro et al. (2021) found that communities at different degrees of extremism exhibit important overlaps, and that users tend to migrate from more moderate to more extreme ones. Kaiser and Rauchfleisch (2020) measured strong homophily (i.e., patterns of homogeneous associations) in networks of channel recommendations, with implication for the formation of far-right communities; they also claim that YouTube's recommendations point from moderate towards extremist channels more often than the other way around. In a study focusing more specifically on conspiracy theories, Alfano et al. (2021) found that following recommendations leads to a substantial proportion of conspiracy content when starting the query from specific topics, potentially associated with conspiracies. They specifically found that starting on topics more commonly associated with conspiracies led to a higher level of eventual recommendation homogeneity.

On the other side of the spectrum, in their analysis of how YouTube's recommender system treats more or less politically extreme channels, Ledwich and Zaitsev (2020) overturn common assumptions by showing how mainstream sources seem to be advantaged, compared to extremist ones, in video-level watch-next recommendations obtained through anonymous accounts. In an attempt to audit Google's announcement of a crackdown on conspiracy and misinformation content, Faddoul, Chaslot, and Farid (2020), also focusing on watch-next recommendations, were largely able to corroborate such claims. However, their results signal a persistent (although, weakened) higher likelihood of a conspiracy video generating a conspiracy recommendation.

As this review suggests, despite the growing cleavage between supporters and deniers of the filter bubble and rabbit hole hypotheses, empirical evidence is still inconclusive and/or contradictory. One reason for this, besides the methodological and epistemological complexity of the issue, might be in the variety of research designs that can be developed to tackle it. The most notable aspect is the tendency of most research on *personalization* algorithms to rely on *non-personalized* recommendations.

Most existing studies discard content-personalization by adopting anonymous users not logged in with an account (Faddoul, Chaslot, and Farid 2020; Ledwich and Zaitsev 2020; Roth, Mazières, and Menezes 2020; Ribeiro et al. 2021) or by using YouTube's API outcome as proxies for recommendations (Airoldi, Beraldo, and Gandini 2016; O'Callaghan et al. 2013). However, actual recommendations are not the deterministic output of pre-existing related video networks, such as those retrievable *via* YouTube API or *via* one-shot anonymous scraping sessions. In order to audit the functioning of a recommender system in real-world scenarios, one needs to engage with personalization (Milan and Agosti 2019), and this can only be achieved by looking at longitudinal data of (real or simulated) users, possibly logged in with their Google Account credentials. By making use of some form of personalization (priming non-logged-in users with an ideologically-polarized watch history and retaining cookies), Romano et al. (2021) provide preliminary evidence for the emergence of ideology-specific recommendation patterns.

As a corollary, most existing studies try to assess the behavior of recommender systems based on one standard click strategy (e.g., the first watch-next recommendation); however, there is no such a thing as an "average user" (Roth, Mazières, and Menezes 2020), and the outcome of algorithms might be more or less sensitive to different users' behaviors. In particular, the tendency of users to make use of (more or less) personalized recommendations as *input* for their watch patterns can have a substantial impact on the videos recommended as *output* (Solsman 2018; Cooper 2020). We can expect that the fact that a YouTube user primarily watches recommended videos might be interpreted as implicit positive feedback, further steering recommendations toward a certain type of content.

Moreover, little is known about how quickly a user's recommendations adapt to a user's behavior, even though this is a critical aspect when it comes to the generation of so-called rabbit holes, as several studies generally focus on one-step recommendations. In other words, we do not know how strong a user's signaled preference for certain types of content needs to be in order for it to effectively steer recommendations towards that type of content, nor how easy or difficult it is to revert the effects of a user's priming on the recommendation system.

Against this background, the present study followed an original experimental setup that simulates actual patterns of personalized recommendations. It adopted sock-puppet Google accounts programmed to follow different watch patterns, each carrying their watch history to the next step of recommendations. This allows us to assess the impact of different watch strategies, all signaling interest towards conspiracy content by following less or more personalized inputs, on the strength and dynamics of the recommendation system's preference for conspiracy videos.

Materials and Methods

In order to determine how different watch strategies affect YouTube personalized recommendations in relation to conspiracy content, a sock-puppet algorithmic audit approach (Sandvig et al. 2014) was designed. Twenty ad hoc Google accounts have been generated, and programmatically controlled to operate on YouTube according to four distinct watch strategies modeling four distinct users' behavior, each taking more and more personalized recommendations as input. The bots, logged in through their Google account, watched 15 videos in sequence each, and at each iteration, the top 20 homepage recommendations were scraped and classified into conspiracy versus non-conspiracy content using a support-vector machine (appendix Machine Learning). Consequently, it was possible to characterize the effects of different watch strategies on the network of video recommendations produced in terms of the proportion of conspiracy content over time. Moreover, another experiment was conducted in order to assess how long would it take for each of the model users "trapped" in a (potential) bubble to "escape" it. The article thus compares the effect of different watch strategies on the tendency for a user to spiral down into a stream of conspiracy content, as well as the persistence of said (potentially) pernicious algorithmic outcome. Data were collected in the first week of May 2021. The following sections describe the experimental setup.

Watching Conspiracy Videos

Google Login

Google's strict policy regarding automated activity poses many obstacles to logging into a Google account using automation software. To circumvent this restriction, two steps had to be taken. Firstly, the Selenium WebDriver[1] was paired with the Selenium-Stealth package,[2] which removes metadata about the current browser, thus masking the fact a WebDriver is being used. Removing metadata causes Google's login service to trigger a warning that prevents a user from logging in. To avoid this warning, the Google accounts were created within the WebDriver. Therefore, all 20 accounts were manually created using ChromeDriver. Since Google accounts require a phone number verification upon creation, ten free (prepaid) SIM cards were ordered from various providers in order to create the accounts. Each SIM card could create two to three accounts before being blocked due to suspicious activity.

The Watch Strategies

After all accounts had been created, they were divided into four distinct watch strategies, each representing distinct user behavior. Each strategy was implemented by 5 bots.

1. *Random non-conspiracies (baseline)*. In the first watch strategy bots watch random *non-conspiracy* videos from a dataset (for a description of the dataset see appendix Dataset). This watch strategy is used as the baseline to compare the other three strategies.
2. *Random conspiracies*. In the second strategy bots watch random *conspiracy* videos from the labeled dataset. This watch strategy should provide signals of interest toward conspiracy content through a watch pattern not based on inputs already subject to personalization.
3. *Watch-next recommendations*. The third strategy relies on the watch-next, partly personalized recommendations as input, although it consists of more stages. It starts as strategy 2 (it chooses a random conspiracy video) after which the bots watch the four most similar videos in the dataset (similarity was defined as the cosine similarity of the titles, descriptions, transcripts, channel descriptions, and channel keywords of the videos) in order to allow for the algorithm to “get a feel” for the user’s interests. After watching those five initial videos, it starts looking at the recommended videos displayed next to the current video and selects the one that is most likely to be a conspiracy video (out of the first 20 recommendations). The watch-next recommendations are based on both the content of the current video and the user’s past behavior.
4. *Homepage recommendations*. The fourth strategy is similar to the previous one, but rather than choosing a recommended conspiracy video from the watch-next recommendations listed next to the current video, it chooses a (likely) conspiracy video from the recommendations listed on the user’s homepage (again, out of the top 20). The homepage recommendations include more personalized content than the watch-next recommendations (Roth, Mazières, and Menezes 2020); compared to the third strategy, this will lead to the user watching more personalized, rather than content-based, recommendations.

For strategies 3 and 4, the likelihood of a recommendation being a conspiracy video was estimated by a neural network (see appendix Machine Learning) using the title, description, transcript, channel description, and channel keywords of the specific video.

Each strategy was executed by 5 different accounts in order to control for random fluctuations. Each account watched a total of 15 videos as described by their watch strategy, for a total of 300 videos. To simulate real-world user behavior, the average watch time proportion for the videos was normally distributed with a mean of 55% and a standard deviation of 25% (Park, Naaman, and Berger 2021; Lang 2018). In the same vein, the clicking behavior of users was simulated as accurately as possible. Whenever, for strategies 3 and 4, none of the recommendations were predicted to be conspiracy videos, the probability of a user clicking on a video at position *k* within a given list of recommendations (its click-through rate: *CTR*), was determined using the following formula:

$$CTR(k;N,\alpha)=\frac{1/k^{a}}{\sum_{n=1}^{N}(1/n^{\alpha})} \tag{1}$$

Wherein *N* is the total number of recommendations (*N* = 20 in our case) and *α* is the distribution's exponent value ($a\times b\times d$) (Zhou, Khemmarat, and Gao 2010). Using this formula, when considering the first twenty recommendations, the first recommendation will have a click-through rate of approximately 20.6%, after which the CTR quickly decreases, until a probability of 1.9% at the twentieth recommendation.

Running the Bots

After the accounts were logged in, they started watching YouTube videos according to their watch strategy. However, some restrictions were implemented to avoid excessive total watch time. For example, bots were not allowed to watch videos over an hour long, nor live streams that can theoretically go on for an indefinite amount of time. Additionally, the random videos at the start of the third and fourth strategies were first manually inspected to make sure the bots would not start the experiment by watching a falsely flagged conspiracy video. Considering the potential presence of false positives in the tested classification outcome, it is possible that a few videos that are flagged as conspiracy videos are in reality not. Despite the likelihood of a false positive being generally low, the selection of a non-conspiracy video as a seed for the watch strategies could have substantially altered the results, hence the manual intervention to avoid the possibility.

Data and Analysis

Running the script for all 20 bots resulted in two different datasets: the first containing the videos watched by the bots (watched videos, 300 in total) and the second containing the homepage recommendations for all bots after each video watched (homepage recommendations, 6,000 in total). The outcome variable has been computed at each step (i.e., number of videos watched) by looking at the top 20 homepage recommendations, rather than on "watch-next," because those appear to be the most personalized and less researched ones (Faddoul, Chaslot, and Farid 2020). These recommendations were then labeled as being either conspiracy or non-conspiracy videos by the classifier. Additional information about the video was collected *via* YouTube's API: title, description, transcript, (dis)likes, views, video duration, channel description, and channel keywords. The title, description, transcript, channel description, and channel keywords were collected in order for the classifier to label the videos. Since the dataset used to prime the watch strategies dates back to 2017, a small number of its videos are not hosted on the platform anymore. Additionally, some videos encountered by the bots were labeled as age restricted by YouTube, prohibiting the bots from watching them. This resulted in a small number of missing observations. In order to avoid further reducing the sample size, we decided to estimate the missing values by averaging the values at the previous and at the following step.

The data collected allowed us to compare the proportion of recommended videos classified as conspiracy content across different watch strategies. Descriptively, this consisted in calculating the ratio of conspiracy-classified recommendations for each "treatment" strategy (strategy 2, 3, and 4) over the baseline (strategy 1). The ratio is calculated for each number *n* of watched videos as the number of the top 20 recommendations classified as conspiracy videos for strategy *S*, divided by the same measure but then for the random non-conspiracy strategy 1. The ratio thus indicates how much more conspiracy videos are recommended in strategy *S* than in the random non-conspiracy watch strategy.

In order to test for statistical significance, we conducted a Mixed ANOVA using the number of videos watched as the within-subject factor, the watch strategy followed by each group of bots as the between-subject factor, and the proportion of conspiracy-classified videos over the 20 collected videos for each step as the dependent variable. This model allowed us to test whether, overall, the number of videos watched and the different watch strategies influence the behavior of the recommendation system. Moreover, we focused on the interaction effects in the pairwise comparisons; this made it possible to determine at which number of videos watched each relevant watch strategy exhibits significant differences from the baseline strategy, understanding the dynamics of "falling" into a conspiracy bubble. In order to unpack the effects of watch strategies on recommendations further, we computed and compared, for each strategy, the conditional probabilities of within-class recommendations (i.e., non-conspiracy content producing non-conspiracy recommendations and conspiracy content producing conspiracy recommendations) and across-class recommendations (i.e., non-conspiracy content producing conspiracy recommendations and conspiracy content producing non-conspiracy recommendations).

Persistence of Recommendation Patterns

Another experiment was set up in order to assess the influence of different watch strategies on the *persistence* of (potential) recommendation patterns generated through the previous one.

Since watch strategy 1 represents the baseline, its five bots were ignored. The remaining bots, still primed with their respective watch strategy, were fed a total of 15 non-conspiracy videos, and their homepage recommendations were stored as in the previous setup. The title, description, transcript, channel description, and channel keywords of each recommendation were again downloaded for the classifier to predict whether each recommendation was a conspiracy video. The proportion of conspiracy-classified recommendations was subsequently analyzed again in terms of watch strategy and the number of videos watched. By doing so, it was possible to observe how many videos each bot needed to watch for its recommendations to start looking similar to that of the baseline again, hence "escaping" a bubble of conspiracy-related content. The Mixed ANOVA post hoc interaction effects between non-baseline strategies versus the baseline at each number of videos watched were used again to test the persistence of the effects created by different watch strategies.

Conspiracy Videos Classification

The whole research relies on the capacity to discriminate between conspiracy and non-conspiracy content. Considering the large number of videos being recommended, determining each video manually would have been unfeasible; hence we trained conspiracy video classifiers based on a dataset containing 7000 YouTube channels manually annotated as a conspiracy channel or not (with 2825 conspiracy channels) (Ledwich and Zaitsev 2020).

For each channel, the title, description, and transcript of the ten most recently uploaded videos were downloaded using YouTube's API. Additionally, the channel description and channel keywords were added to each video. Each video received the conspiracy label of its channel. Table A.1 contains example data. This dataset contained 65.683 unique YouTube videos, with 22.156 labeled as conspiracy. The two classes were balanced by undersampling non-conspiracy videos so that the classifier would not develop a bias for non-conspiracy videos (Lemaître, Nogueira, and Aridas 2017; Sun, Kamel, and Wang 2006). Considering the large size of the dataset, undersampling was preferred over implementing class-weights (Brownlee 2021). The resulting set contained 44.312 videos.

Performance Optimization

After splitting the dataset into a training (80% of data), validation (10% of data), and test (10% of data) set, the hyperparameters of five different algorithms were tuned to get the optimal performance (Feurer and Hutter 2019). Performance was measured using four distinct metrics: accuracy, which shows the share of correct predictions; recall, which shows what fraction of truly positive samples were correctly labeled as such; precision, which shows what part of the positive predictions were correct; and the F1-score, which is the harmonic mean of the recall and precision (Sokolova and Lapalme 2009). For each classifier, different configurations of hyperparameters (such as the kernel and the penalty parameter) were systemically tested – each possible combination was tried. The classifiers were trained on the training set and the optimal hyperparameters were determined based on the performance of the classifiers on the validation set. By saving these performance measures for every configuration, for every classifier, the optimal configuration of each classifier could be determined. Lastly, the classifiers were equipped with their optimal hyperparameters and then tested for on the test set. By comparing the performance of every optimally configured classifier on the test set, the best-performing classifier could be chosen (Reitermanova 2010).

We trained video classifiers using five approaches (k-nearest neighbors, support-vector machine, neural network, logistic regression, and ridge regression) and an ensemble. Apart from k-nearest neighbors, all classifiers scored very similarly with F1 scores between .9 and .92, see Figure 1. Appendix Machine Learning contains the details on the hyperparameter settings used for each method.

We applied the classifier to new unseen data encountered in the experiments. We used the SVM for the final counts and the neural network for the real-time classification of recommendations because of its inherent use of probabilities, rather than confidence scores.

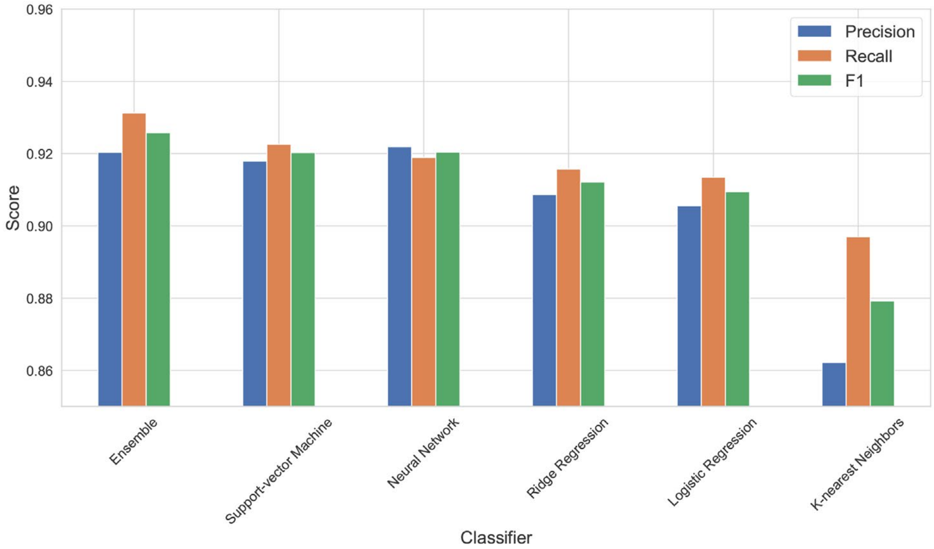


Figure 1. Precison, Recall, and F1-scores for video conspiracy classification for each classifier with optimized hyperparameters (average score over both classes). A train-validation-test split was performed to determine model performance. Classifiers were trained on 80% of the data ($N = 35.450$). Optimal hyperparameters were determined based on performance on the validation set ($N = 4.431$, 10% of the data). The shown scores are the performance of the optimized models on the test set, which consisted of the left 10% of the data.

Although the classifier's outcome cannot be considered 100% accurate, and the notion of "conspiracy-related content" can be hard to define even by human coders, inaccuracies or ambiguities in the classification task do not significantly undermine the main goal of the experiments: that of assessing the respective effects of different watch patterns on the behavior of YouTube's recommendation system.

Results

This section presents the results of our two experiments, answering our research questions.

Experiment 1: Spiraling into Conspiracy Bubbles

Aggregate Ratios

The first experiment measured the tendency of YouTube's recommender to develop a preference for conspiracy videos depending on the different watch strategies. Table 1 presents the ratio between the total number of recommended conspiracy content, accumulated over 15 videos watched by 5 bots per strategy, versus the non-conspiracy baseline (strategy 1). This value can be interpreted as the increase in the likelihood of YouTube's recommender suggesting conspiracy content. These results show that watch patterns based on conspiracy content do produce a largely higher proportion of conspiracy content as recommendations. Moreover, the effect is stronger for watch strategies involving personalized content.

Effects of Strategies and Videos Watched

The results presented refer to aggregated values over a sample of observations (5 bots) and a number of steps (15 videos watched) per watch strategy. In order to test if the differences are statistically significant, we ran a Mixed ANOVA model, as described in Materials and methods (Table 2).

The watch strategies have significant effects on the proportion of conspiracy-classified recommendations ($_t$). The number of videos watched has a (positive) significant influence on the proportion of conspiracy content recommended $_l$). There are also significant interaction effects of watch strategies and the number of videos watched ($_w$), signaling that the effects of different watch strategies are significantly different in different points of "time" (i.e., videos watched). Thus, the different strategies have a significant impact on the number of conspiracy videos being recommended, as well as the "speed" with which this impact occurs.

As our main goal is to assess the difference between treatment strategies and the baseline, we turn to the post hoc pairwise comparisons to evaluate differences between specific pairs of strategies (Table 3).

More or Less Personalized Strategies

Having assessed that watch patterns including conspiracy content tend to generate significantly more conspiracy recommendations, we now turn to test whether increasingly personalized watch strategies tend to have stronger effects on this pattern. This hypothesis seems corroborated by the descriptive aggregate and average values presented in Table 1 and Figure 2. In order to do so, we look at the post hoc comparisons contrasting individual treatment strategies reported in Table 3. Whereas the odds ratio effects reported before are pretty clear in indicating a hierarchy between watch strategies in terms of "degree of personalization," the results of the statistical test are more ambiguous. As Table 3 shows, all the treatment strategies significantly differ from the baseline (strategy 1) in terms of the proportion of conspiracy-classified content recommended to the bots. As already noted in terms of ratios, and reported here in terms of the difference in means of probabilities (MD), the effect seems to be stronger the more personalized the watch strategy is (e.g., the difference between strategy 4 and strategy 1 is 35.5%, while the difference between strategy 2 and strategy 1 is 18.7%). However, when confronting different treatment strategies, the only statistically significant result is between strategy 4 (fully personalized), and strategy 2 (non-personalized). Despite a remarkable difference in terms of ratios (see 1), no statistically significant difference is observable between strategy 3 (partly personalized) and strategy 2, and between strategies 4 and 3. This ambiguity is also evident in Figure 2 when looking at the large intersection in confidence intervals, represented by the overlap in the areas around the lines.

Table 1. Ratio of total conspiracy-classified recommendations for each strategy versus the baseline (for each strategy averaged over 5 users).

	Random conspiracy	Watch-next	Homepage
Ratio vs. strategy 1	3.8	4.8	6.3

E.g., after watching 15 videos following strategy 2, a user gets recommended 3.8 times more conspiracy videos than when watching according to the baseline watch strategy.

Table 2. Results of the Mixed ANOVA model with watch strategy (between factor), number of videos watched (within factor), and proportion of conspiracy-classified content (dependent variables).

Factor	DF1	DF2	*F*	*p*	x_a
strategy	3	16	18.47	0.000	.41
vids_watched	14	224	12.78	0.000	.16
interaction	42	224	2.69	0.000	.10

Table 3. Post hoc pairwise comparison of Mixed ANOVA model with watch strategy (between factor), number of videos watched (within factor), and proportion of conspiracy-classified content (dependent variables).

A	B	*p*	MD
Baseline	Random conspiracy	0.000	−0.187
Baseline	Watch-next	0.002	−0.251
Baseline	Homepage	0.003	−0.355
Random conspiracy	Watch-next	0.160	−0.064
Random conspiracy	Homepage	0.041	−0.168
Watch-next	Homepage	0.168	−0.105

Strategies' Dynamics versus the Baseline

Figure 2 plots the proportion of conspiracy-classified content at each moment in time for the four watch strategies, thus showing the evolution of conspiracy-related "bubbles" per each watch strategy.

As Figure 2 indicates, all strategies (excluding the baseline) quickly (after 4 to 6 initial videos) lead to a permanent significant increase in the number of conspiracy recommendations on the user's homepage. This figure gives a more fine-grained picture of the experiment compared to only the end results given in Table 1.

Strategy 2 (random conspiracy videos) produces the lower overall effects, and takes the longest to generate consistently significant differences from the baseline. Besides an impromptu significant difference at step 1, bots following this strategy need to watch 6 videos before the difference with the baseline becomes stably significant, with 21.5% of its recommendations consisting of conspiracy-labeled videos (3.1 times higher than the baseline). The proportion of conspiracy videos being recommended to the users of this strategy keeps steadily growing, eventually reaching its maximum at 42%, a measure 5.3 higher than the corresponding value in the baseline.

Strategy 3 (watch-next recommendations) spiraled towards a conspiracy bubble quicker, with 4 videos needed for the difference with the baseline to become significant at 26% (corresponding to a ratio of 4.7). Following this head start, the increase comes to a halt, with step 7 producing non-significant differences, and declining after peaking at 9 videos watched with 46% of conspiracy content recommended (corresponding to a ratio over the baseline of 6.6). The percentage of conspiracy recommendations settles then around the value of 35%. Overall, compared to strategy 2 (random conspiracy), strategy 3 presents a higher proportion of conspiracy-related recommendations until a steady decline from video 10.

Strategy 4 (homepage recommendations) also takes 4 steps to generate significant differences (with 23%, 4.2 times higher than the baseline), however, it ends up being the watch strategy that causes, significantly, the most conspiracy

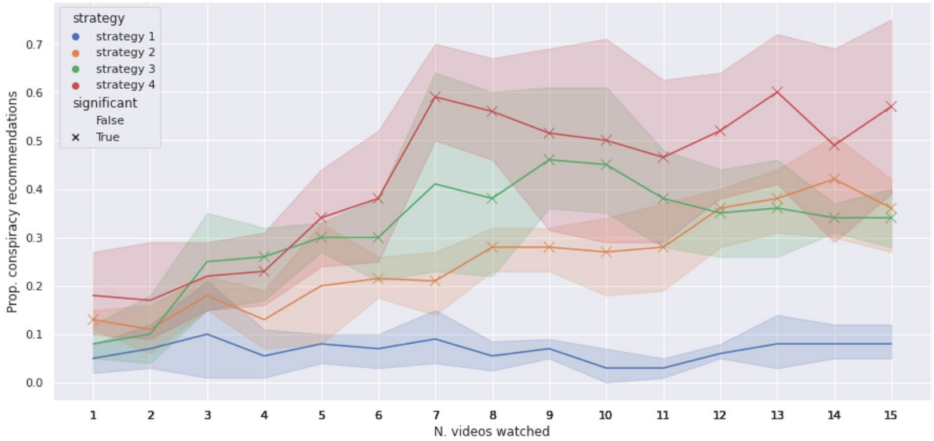


Figure 2. Proportion of conspiracy recommendations after each number of videos watched per strategy. Line values correspond to averages, areas correspond to 95% confidence intervals. Crossed dots correspond to significant differences ($p < 0.05$) compared to the baseline (strategy 1). *Note:* Significance to the baseline was tested through independent samples *t*-tests between each strategy and the baseline at each number of videos watched.

content being recommended. By definition, strategy 3 (watch-next) and strategy 4 (homepage) behave identically until video number 5, and indeed differences between the two strategies up to then are not observable. However, as soon as strategy 4 diverges from strategy 3 by taking into account homepage recommendations, the percentage of conspiracy recommendations increases drastically, quickly reaching 58% at video watched number 7 (corresponding to a ratio over the baseline of 6.6). After this, the value oscillates between 46% and 60% until the end of the experiment, producing at step 10 almost 17 times more conspiracy-classified recommendations than the baseline. These results indicate that, in terms of average values, the homepage recommendations seem to have stronger effects on overall recommendations.

Experiment 2: Escaping from the Spiral

After assessing the relative consistency of conspiracy-related recommendations, and the pace at which different watch strategies fall within this pattern, it is interesting to measure how long it takes for the recommender system to stop recommending conspiracy content at an above-baseline rate. If the former can be considered a proxy of whether and how easily conspiracy bubbles are generated, the latter tells us about the persistence of those recommendation patterns. Each of the bots was subsequently fed 15 non-conspiracy videos, and their recommendations were collected at each step, to measure the impact of this “de-radicalization” strategy. Figure 3 shows the dynamics of the percentage of conspiracy-related videos over the number of videos watched, grouped per watch strategy. Again, the statistical significance in differences with the baseline at each step is indicated based on the post hoc comparison following the same Mixed ANOVA model, using strategy 1’s data from the previous experiment and this experiment’s data for the treatment strategies.

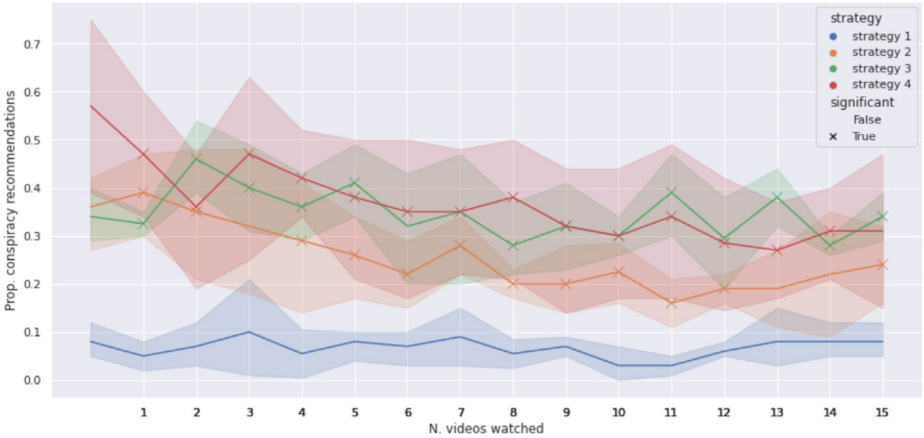


Figure 3. Proportion of conspiracy recommendations after each number of non-conspiracy videos watched per strategy (primed in experiment 1). Line values correspond to averages, areas correspond to 95% confidence intervals. Crossed dots correspond to significant differences ($p < 0.05$) compared to the baseline (strategy 1).

While in the previous experiment it took only a handful of watched videos for treatment strategies' recommendations to start preferring conspiracy content, escaping this "spiral" seems more difficult. In all relevant watch strategies, the share of conspiracy content recommended manifests a slow tendency to decrease, and the difference with the baselines sporadically chases to be significant, especially for strategy 2 (non-personalized). However, the value remains significantly higher than the baseline in most of the points of observations, even after the users have watched a substantial number (up to 15) of non-conspiracy videos.

Discussion

Interpretation of the Results

RQ1 – Proportion of Conspiracy Recommendations

The results of our experiment provide strong evidence that YouTube's homepage recommendation system is highly sensitive to the signals provided by specific watch patterns. When primed with videos classified as conspiracy content, the probability of further conspiracy-classified content being recommended increases dramatically for each of the three "treatment" watch strategies developed. Moreover, within each watch strategy, the number of videos watched produces a significant effect on the proportion of conspiracy-classified recommendations, indicating that these effects are cumulative. The dynamic of recommendations across the number of videos watched systematically generates, relatively early, significant differences between treatment and baseline watch strategies at different numbers of videos watched. Watching random non-conspiracy videos does not produce any trend in the proportion of conspiracy videos recommended. Watching conspiracy content, instead, quickly introduces a preference for further conspiracy content.

Overall, these results suggest that YouTube's homepage recommendation system is prone to considerably, systematically, and quickly recommending a disproportionally higher percentage of certain types of content if the user signals interest towards this type of content through their watching behavior – even when it comes to potentially problematic content, such as conspiracies.

RQ2 – Reverting the Pattern

As the results of the second experiment indicate, reverting this pattern proves more difficult than generating it. Despite the proportion of conspiracy content somewhat declining once users stop signaling interest through their watch patterns, even after 15 unrelated videos the algorithm is in most cases still inclined to recommend a significantly higher proportion of conspiracy content than the baseline. The observed trend does not seem to suggest that a further drop would be only few steps ahead. Moreover, the decline in preference for conspiracy recommendations seems moderate, especially for watch strategies adopting personalized content as input (strategies 3 and 4).

RQ3 – Personalized Strategies

Different treatment strategies have been designed according to different logic: strategy 2 (random conspiracies) is based on non-personalized inputs; strategy 3 (watch-next recommendations) is based on partly personalized inputs; and strategy 4 (homepage recommendations) is based on fully personalized inputs. Whereas the effect of treatment strategies on recommendation patterns is evident both in terms of descriptive ratios and in terms of statistical tests, the results are more ambiguous when it comes to differences between specific watch strategies, i.e., the role of personalized inputs in influencing the recommendation output. Based on aggregate results, the more personalized the watch strategy, the higher the ratio of conspiracy recommendations versus the baseline, as well as the higher the probability of more conspiracy content being recommended as output to conspiracy content being watched as input. This observation holds true in terms of average proportions of conspiracy-classified recommendations over the number of videos watched: differences in average proportions at each step seem to be consistent with the idea of the stronger effect of more personalized strategies. However, the only difference that produces statistically significant results is the one between strategy 4 (fully personalized) and strategy 2 (non-personalized). Consequently, we cannot draw definite conclusions on the hypothesis that the more personalized the input of the watch strategies, the more personalized the output of the recommender system, although the overall average measures and the comparison between the fully personalized and the non-personalized strategies provide indications in this sense.

Rabbit Holes and Filter Bubbles vs. Algorithmic Personalization

In formulating our research questions and interpreting our results, we avoided making bold and specific references to notions such as “rabbit holes,” “filter bubbles,” and the fallacies of YouTube’s content moderation, as these are generally hard to define in operative terms. Nonetheless, our article contributes clear results and an innovative methodology to these tangent debates. Our results are compatible with the idea that YouTube's recommender system is susceptible to the creation of filter bubbles and/or rabbit holes even when it comes to debatable, potentially harmful content such as conspiracy videos. Despite Google's efforts to reduce the visibility of such content in its recommendations (YouTube 2019), a real-world scenario of watch patterns easily, quickly, and persistently generates a substantial preference for conspiracy content based on signals of interest. This effect might be particularly strong when users watch videos recommended by the platform, either as watch-next or (especially) homepage recommendations, although more evidence is needed in this sense.

Existing studies on YouTube's propensity to generate filter bubbles in general, or filter bubbles of extremist content in particular, thus far provided contrasting evidence. In spite of the growing skepticism around the role of recommendation systems in trapping users in bubbles or rabbit holes of extremist content (Bruns 2019; Hosseinmardi et al. 2020; Zuiderveen Borgesius et al. 2016), Roth, Mazières, and Menezes (2020) highlighted a tendency for recommendations to generate bubbles of homogeneous content in terms of broad topics, and Romano et al. (2021) provided support to the argument that YouTube's recommendation system is sensitive to personalized input in terms of sources and content recommended along political partisanship lines. The idea of extremist (mostly, right-wing) "rabbit holes" being a worrisome trait of the platform's algorithmic system (Tufekci 2018) is supported by some studies (O'Callaghan et al. 2013; Ribeiro et al. 2021) and disconfirmed by others (Hosseinmardi et al. 2020; Ledwich and Zaitsev 2020). Focusing on conspiracy content more specifically, Faddoul, Chaslot, and Farid (2020) were able to back Google's claims about the reduced visibility of such content after the 2019 crackdown, however, they also highlight that conspiracy videos are still likely to generate more conspiracy videos in watch-next recommendations. Such an inconclusive body of evidence is surely related to the infancy of the field of study and the number of factors at play, as well as the heterogeneity in research designs adopted.

Whereas, mostly for reasons of convenience, most existing studies discard personalization (i.e., users' watch history stored in the browser or, even better, in Google Accounts) from their analysis, this is a crucial aspect to consider when assessing the concrete effects of algorithms such as recommender systems on users' information diets (Milan and Agosti 2019). The present study addresses *personalization* on three levels. First (and foremost), it deals with personalized (albeit fictional) recommendations as the outcome variable, instead of observing anonymous users or API responses. Using bots logged in with fresh-made Google accounts allowed our experimental setup to better approximate the natural conditions of recommendation patterns involving real-world users. Recommendation systems such as YouTube's are complex algorithms aggregating inputs related to content and to users' behavior. They are dynamic and local systems that, when reduced to static and universal proxies, generate misleading, ecologically invalid operational definitions of what "algorithmic recommendations" are. Secondly, this research distinguishes between personalized and non-personalized recommendations as input. This permits us to reflect upon how different sources that bring a user to watch a certain video (i.e., external referral, videos suggested after the current one, or videos displayed on the homepage) might have different impacts on the variety and quality of the content they are recommended. By differentiating watch strategies, this article shows that a user's individual watching behavior influences the strength of recommendation patterns as well as their persistence. Following recommendations as input might generate stronger effects on recommendation patterns than accessing videos through their URLs, and following homepage recommendations might generate different results than following watch-next ones. Thirdly, while most existing research focuses on the video-specific recommendations listed next to the video currently being watched, generated also based on information about the video itself, this research collected homepage recommendations, solely based on the analysis of personalized watch patterns. Whereas we do not know the relative impact of these recommendations on users' watching behavior, the fact that, based on signals of interest, the homepage of a YouTube user gets so quickly populated with problematic content might be considered an issue of societal concern.

Our results do not provide definite proof of the existence of such things as filter bubbles or rabbit holes as outcomes of YouTube's recommender system. However, they contribute clear evidence in this direction based on a more ecologically valid methodological framework than that of most existing studies. We refrained from providing specific conceptual and operative definitions of notions such as filter bubbles and rabbit holes because our data can still be open to interpretation based on methodological, epistemological as well as normative assumptions. However, by tackling the issue of personalization at many levels, our study contributes to this developing debate by providing evidence for the potentially pernicious effects of YouTube's recommender system, as well as a replicable methodology to investigate this issue further.

Limitations and Expansion

An important limitation to acknowledge relates to the small sample size. The key aspect of our methodology is that, unlike existing studies, the recommender system is observed in response to authenticated users carrying a watch history. While this requires additional set-up costs, the ability to research algorithmic effects in a quasi-realistic scenario arguably outweighs the drawbacks. However, due to the time-demanding requirements associated with setting up ad hoc Google accounts and having them watch substantial portions of videos make this method difficult (i.e., expensive) to replicate on a large scale. As a result, a limitation of this research is the fairly small sample size (5 user-level observations per group). Despite the small sample size, the effects observed in terms of treatment strategies versus the baseline are large enough to hold. The small sample size, paired with a relatively high variance in the distributions of, especially, strategy 3 and strategy 4, probably explains the ambiguous results obtained when comparing the effects of specific treatment strategies. Future research with more resources available could replicate the present study on a larger scale. Parallelization of execution could be a technical solution to the issue of scale. Another possibility would be that of relaxing the experimental character of the research design, and involving a number of volunteers in running collective tests on a large scale (WeTest[3]), either with ad hoc accounts or with their actual ones, mostly primed by a realistic, long-term watch history.

Users' watch strategies and the computation of the outcome variable both required YouTube's content to be automatically classified as conspiracy or non-conspiracy. We are aware of the fact that such automated classification is prone to errors and presents a general shade of ambiguity. Quantifying and correcting for this distortion can be quite a daunting task, considering that assessing what constitutes conspiracy-related content can be an inherently ambiguous operation for human coders as well. However, evaluating and coping with the incidence of content misclassification is not a priority for the goal of our research, as we focused on diverging trends between treatment and baseline strategies, and we have no reasons to believe that any systematic error could be introduced by different watch strategies.

Whereas adopting logged-in accounts with different watch strategies provide a more realistic and more meaningful way to study the effects on YouTube's recommendations system, the requirements of an experimental setup maintain elements of artificiality in the users' profiles and behaviors, affecting ecological validity. Relying on fresh-made accounts with a straightforward watch history dictated by a relatively deterministic watch strategy might introduce distortions in the results when compared with existing, long-term accounts with a more complex watch history (and other web browsing activities). This is particularly true for the dynamic of patterns, investigated in this study, as the lack of alternative potential "interests" signaled by the users might explain the quickness and persistence of the patterns observed. Tests involving real users, as suggested above, might help control for this potential distortion. Furthermore, although necessary for the experimental design, the accounts' behavior of constantly returning to the YouTube homepage before watching the next video, is unnatural. Although there are no indications that lead us to believe this influenced the results, it is good to keep this atypical behavior in mind.

Arguably, assessing the relevance of issues such as filter bubbles and rabbit holes requires accounting for actual, "naturalistic" patterns of information consumption, rather than on recommendations generated and selected based on artificial settings (Hosseinmardi et al. 2020). Nonetheless, in the spirit of auditing algorithms of societal relevance (Sandvig et al. 2014), the existence of concerning algorithmic outcomes on the basis of artificial set-ups is an issue worth raising and assessing.

Conclusion

The goal of this research was to find out how YouTube's recommendation system treats conspiracy-classified content based on personalized watch patterns. By automating 20 brand-new YouTube accounts, we set up 4 different watch strategies to prime users with either non-conspiracy or conspiracy-classified content, based on either non-personalized or personalized inputs. Our results show that watching conspiracy-classified content produces a strong and significant increment in the proportion of conspiracy-classified content being recommended. Reverting this pattern, by feeding primed users non-conspiracy-classified content, proves substantially more difficult than generating it. Watch strategies that rely on the platform's own recommendations have the largest effect.

On a general level, although our results do not definitively prove filter bubbles or rabbit holes are a result of YouTube's recommendation algorithm, they do provide evidence in this direction based on a more holistic and sound methodological framework. While this evidence is far from conclusive, this study contributes clear findings to the debate based on an original experimental setting.

Previous research on the topic, coming to conflicting conclusions, has generally discarded personalization from their research design. This research, instead, engaged with personalization by using logged-in accounts, by differentiating between more or less personalized watch strategies, and by focusing on the homepage, purely personal recommendations. The variety of operational choices related to assessing conceptually malleable concepts such as filter bubbles or rabbit holes (what type of recommendations to collect? what type of users to model? what type of topics to focus on? what type of patterns to look for?) hamper the possibility to produce cumulative knowledge on the issue. Converging on a more solid, coherent body of results, we argue, requires a first step to take personalization seriously when studying recommendation systems.

Notes

1. https://pypi.org/project/selenium/
2. https://pypi.org/project/selenium-stealth/
3. For an example of such research design see https://youtube.tracking.exposed/wetest/1/

Acknowledgments

This research was supported in part by the Netherlands Organization for Scientific Research through the ACCESS project grant CISC.CC.016 (https://www.nwo.nl/en/projects/ciscc016). Furthermore, we thank Emma Tas for her useful discussions and continuous feedback, and the anonymous referees for their extensive and helpful reviews.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This work was funded by Nederlandse Organisatie voor Wetenschappelijk Onderzoek.

ORCID

Roan Schellingerhout http://orcid.org/http://orcid.org/0000-0002-7388-309X

References

Airoldi, Massimo, Davide Beraldo, and Alessandro Gandini. 2016. “Follow the Algorithm: An Exploratory Investigation of Music on YouTube.” *Poetics* 57: 1–13.

Alfano, Mark, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. 2021. “Technologically Scaffolded Atypical Cognition: The Case of YouTube’s Recommender System.” *Synthese* 199 (1–2): 835–858.

Allington, Daniel, and Tanvi Joshi. 2020. “What Others Dare Not Say’: An Antisemitic Conspiracy Fantasy and Its YouTube Audience.” *Journal of Contemporary Antisemitism* 3 (1): 35–54.

Beraldo, Davide, and Stefania Milan. 2019. “From Data Politics to the Contentious Politics of Data.” *Big Data & Society* 6 (2): 205395171988596.

Birch, Matthew K. S. 2019. “White Rabbit. The Logic and Proportion of Conspiracy Theory Videos on YouTube: A Foucauldian Discourse Analysis.” Master’s thesis. Malmö universitet/Kultur och samhälle.

Bozdag, E., and J. van den Hoven. 2015. “Breaking the Filter Bubble: Democracy and Design.” *Ethics and Information Technology* 17 (4): 249–265. DOI: 10.1007/s10676-015-9380-y.

Brownlee, Jason. 2021. “Random Oversampling and Undersampling for Imbalanced Classification.” January. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

Bruns, Axel. 2019. “Filter Bubble.” *Internet Policy Review* 8 (4): 34–48.

Bryant, Lauren V. 2020. “The YouTube Algorithm and the Alt-Right Filter Bubble.” *Open Information Science* 4 (1): 85–90. DOI: 10.1515/opis-2020-0007.

Burges, Chris J. C., and Bernhard Schölkopf. 1997. “Improving the Accuracy and Speed of Support Vector Machines.” *Advances in Neural Information Processing Systems* 9 (1997): 375–381.

Chitra, Uthsav, and Christopher Musco. 2020. “Analyzing the Impact of Filter Bubbles on Social Network Polarization.” In Proceedings of the 13th International Conference on Web Search and Data Mining, 115–123.

Cooper, Paige. 2020. “How Does the YouTube Algorithm Work? A Guide to Getting More Views.” August. https://blog.hootsuite.com/how-the-youtube-algorithm-works/

Covington, Paul, Jay Adams, and Emre Sargin. 2016. “Deep Neural Networks for YouTube Recommendations.” In Proceedings of the 10th ACM Conference on Recommender Systems (RecSys ‘16). Association “For Computing Machinery, New York, NY, USA, 191–198.

Donzelli, Gabriele, Giacomo Palomba, Ileana Federigi, Francesco Aquino, Lorenzo Cioni, Marco Verani, Annalaura Carducci, and Pierluigi Lopalco. 2018. “Misinformation on Vaccination: A Quantitative Analysis of YouTube Videos.” *Human Vaccines & Immunotherapeutics* 14 (7): 1654–1659. DOI: 10.1080/21645515.2018.1454572.

Faddoul, Marc, Guillaume Chaslot, and Hany Farid. 2020. “A Longitudinal Analysis of YouTube’s Promotion of Conspiracy Videos.” CoRR Abs/2003.03318, March. https://arxiv.org/abs/2003.03318.

Feurer, Matthias, and Frank Hutter. 2019. *Hyperparameter Optimization*. Cham: Springer International Publishing, 3–33. DOI: 10.1007/978-3-030-05318-5_1

Freeman, Daniel, and Richard P. Bentall. 2017. “The Concomitants of Conspiracy Concerns.” *Social Psychiatry and Psychiatric Epidemiology* 52 (5): 595–604.

Frey, Dieter. 1986. “Recent Research on Selective Exposure to Information.” *Advances in Experimental Social Psychology* 19 (1986): 41–80.

Frischlich, Lena, Jens H. Hellmann, Felix Brinkschulte, Martin Becker, and Mitja D. Back. 2021. “Right-Wing Authoritarianism, Conspiracy Mentality, and Susceptibility to Distorted Alternative News.” *Social Influence* 16 (1): 24–64.

Furnham, Adrian, and Simmy Grover. 2022. “Do You Have to Be Mad to Believe in Conspiracy Theories? Personality Disorders and Conspiracy Theories.” *The International Journal of Social Psychiatry* 68 (7): 1454–1461. 2022

Gillespie, Tarleton. 2014. *The Relevance of Algorithms* (edited by Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot). Cambridge, MA: MIT Press. https://www.microsoft.com/en-us/research/publication/the-relevance-of-algorithms/

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. “Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance.” *Big Data & Society* 7 (1): 205395171989794.

Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, David M. Rothschild, Markus Mobius, and Duncan J. Watts. 2020. “Evaluating the Scale, Growth, and Origins of Right-Wing Echo Chambers on YouTube.” CoRR Abs/2011.12843, November. https://arxiv.org/abs/2011.12843.

Kaiser, Jonas, and Adrian Rauchfleisch. 2020. “Birds of a Feather Get Recommended Together: Algorithmic Homophily in YouTube’s Channel Recommendations in the United States and Germany.” *Social Media + Society* 6 (4): 205630512096991. DOI: 10.1177/2056305120969914.

Kamarthi, Sagar V., and Stedan Pittner. 1999. “Accelerating Neural Network Training Using Weight Extrapolations.” *Neural Networks* 12 (9): 1285–1299. DOI: 10.1016/S0893-6080(99)00072-6.

Lang, Peter. 2018. "Youtube Average View Duration – The 50% Rule." November. https://uhurunetwork.com/the-50-rule-for-youtube/

Ledwich, Mark, and Anna Zaitsev. 2020. "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization." *First Monday* 25: 3. DOI: 10.5210/fm.v25i3.10419.

Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas. 2017. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *The Journal of Machine Learning Research* 18 (1): 559–563.

Maack, Már Másson. 2019. "YouTube Recommendations Are Toxic," Says Dev Who Worked on the Algorithm, June. https://thenextweb.com/news/youtube-recommendations-toxic-algorithm-google-ai

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27 (1): 415–444. 2001

Milan, Stefanija, and C. Agosti. 2019. "Personalisation Algorithms and Elections: Breaking Free of the Filter Bubble."

Miller, Daniel T. 2021. "Characterizing QAnon: Analysis of YouTube Comments Presents New Conclusions about a Popular Conservative Conspiracy." *First Monday* 26 (2): n. pag. DOI: 10.5210/fm.v26i2.10168.

Neufeld, Dorothy. 2021. "The 50 Most Visited Websites in the World." January. https://www.visualcapitalist.com/the-50-most-visited-websites-in-the-world/

Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2 (2): 175–220. 1998

Noble, Safiya Umoja. 2018. *Algorithms of Oppression*. New York, USA: New York University Press.

O'Callaghan, Derek, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. 2013. "The Extreme Right Filter Bubble." CoRR Abs/1308.6149, August. http://arxiv.org/abs/1308.6149.

Paolillo, John C. 2018. "The Flat Earth Phenomenon on YouTube." *First Monday* 23: 12. DOI: 10.5210/fm.v23i12.8251.

Pariser, Eli. 2011. *The Filter Bubble: What the Internet is Hiding from You*. Penguin UK, London, England.

Park, Minsu, Mor Naaman, and Jonah Berger. 2021. "A Data-Driven Study of View Duration on YouTube." *Proceedings of the International AAAI Conference on Web and Social Media* 10 (1): 651–654. https://ojs.aaai.org/index.php/ICWSM/article/view/14781.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard, MA: Harvard University Press.

Reitermanova, Zuzana. 2010. "Data Splitting." In *WDS'10 Proceedings of Contributed Papers Part I – Mathematics and Computer Sciences*, Vol. 10. Matfyzpress, Prague, Czech Republic, 31–36.

Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2021. "Auditing Radicalization Pathways on YouTube." arXiv: Bib1.

Romano, Salvatore, Davide Beraldo, Giovanni Rossetti, and Leonardo Sanna. 2021. "FilterTube: Investigating Echo Chambers, Filter Bubbles and Polarization on YouTube." https://youtube.tracking.exposed/filtertube/

Rosenbaum, Lisa. 2021. "Escaping Catch-22–Overcoming Covid Vaccine Hesitancy."

Roth, Camille, Antoine Mazières, and Telmo Menezes. 2020. "Tubes and Bubbles Topological Confinement of YouTube Recommendations." *PLoS One* 15 (4): e0231703.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. 2014. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms."

Sokolova, Marina, and Guy Lapalme. 2009. "A Systematic Analysis of Performance Measures for Classification Tasks." *Information Processing & Management* 45 (4): 427–437. 2009

Solsman, Joan E. 2018. "Ever Get Caught in an Unexpected Hourlong YouTube Binge? Thank YouTube AI for That." January. https://www.cnet.com/news/youtube-ces-2018-neal-mohan/

Spohr, Dominic. 2017. "Fake News and Ideological Polarization: Filter Bubbles and Selective Exposure on Social Media." *Business Information Review* 34 (3): 150–160. DOI: 10.1177/0266382117722446.

Sun, Yanmin, Mohamed Kamel, and Yang Wang. 2006. "Boosting for Learning Multiple Classes with Imbalanced Class Distribution." In *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, Hong Kong, China. DOI: 10.1109/icdm.2006.29

Tufekci, Zeynep. 2018. "YouTube, the Great Radicalizer." *The New York Times* 10 (2018): 23–28.

Wahiba, Ben, and Abdessalem Karaa. 2013. “A New Stemmer to Improve Information Retrieval.” *International Journal of Network Security & Its Applications* 5 (4): 143.

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. “The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions.” In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200.

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

YouTube. 2021. “YouTube Community Guidelines & Policies – How YouTube Works.” https://www.youtube.com/howyoutubeworks/policies/community-guidelines/

Zeng, Jing, and Mike S. Schäfer. 2021. “Conceptualizing “Dark Platforms”. Covid-19-Related Conspiracy Theories on 8kun and Gab.” *Digital Journalism* 9 (9): 1321–1343.

Zhou, Renjie, Samamon Khemmarat, and Lixin Gao. 2010. “The Impact of YouTube Recommendation System on Video Views.” In Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC’10). Association for Computing Machinery, New York, NY, USA, 404–410.

Zuiderveen Borgesius, Frederik, Damian Trilling, Judith Möller, Balázs Bodó, Claes H. De Vreese, and Natali Helberger. 2016. “Should we Worry about Filter Bubbles?” *Internet Policy Review* 5 (1): 34–50.

Appendix

Q9

Algorithms All of the data and code used for this research is publicly available on GitHub.

The following script was created and run for all twenty bots, keeping track of the videos they watched and the homepage recommendations they had after each video:

Algorithm 1: Watch YouTube videos according to a watch strategy

```
Data: User information and a video dataset
Result: The watched videos and homepage recommendations of the user

for twenty bots do
    initialize WebDriver;
    log into Google account;
    for fifteen videos do
    if there is a recommendation to be watched then go to the link;
    else
        pick a random video to watch based on usertype;
        determine how long it will get watched;
        go to the link;
    get video metadata and store for overview of watched videos;
    watch video for given amount of time;
    if usertype == 3 then
        pick recommendation next to current video to watch next;
        determine watch time for found recommendation;
    go to YouTube homepage;
    store current recommendations for overview;
    if usertype == 4 then
        pick homepage recommendation to watch next;
        determine watch time for found recommendation;
return watched videos and homepage recommendations;
```

For the second experiment, a slightly altered version of the algorithm was used:

Algorithm 2: Getting out of a filter bubble

```
Data: User information and a video dataset
Result: The watched videos and homepage recommendations of the user

for fifteen bots do
    initialize WebDriver;
    log into Google account;
    for fifteen videos do
        pick a random non-conspiracy video to watch;
        determine how long it will get watched;
        go to the link;
        get video metadata and store for overview of watched videos;
        watch video for given amount of time;
        go to YouTube homepage;
        store current recommendations for overview;
return watched videos and homepage recommendations;
```

Dataset

Data Gathering

To answer the research question, it was necessary to determine which YouTube videos can be considered conspiracy videos. Considering the large number of videos getting recommended, determining each video manually is simply not possible. There are two possible ways to solve this problem. Firstly, there is a dataset that contains nearly 7000 YouTube channels that have been manually labeled based on their political view – almost 3000 of which were labeled as conspiracy channels (Ledwich and Zaitsev 2020); whenever a video is made by one such channel, it can be considered a conspiracy video. Thus, our definition of conspiracy content is any content uploaded by a conspiracy channel, which, according to Ledwich & Zaitsev is: a channel that regularly promotes a variety of conspiracy theories. A conspiracy theory explains an event/circumstance as the result of a secret plot that is not widely accepted to be true (even though sometimes it is). Example conspiracy theories: Moon landings were faked, QAnon & Pizzagate, and Trump colluding with Russia to win the election.

However, due to the enormous amount of existing YouTube channels, the odds of a video being uploaded by a channel that is not present in this dataset are very large. For those videos, a supervised machine learning classifier was used. To optimize performance, five different classifiers have been trained and compared: k-nearest neighbors, support-vector machine, neural network, logistic regression, and ridge regression.

In order to train these machine learning algorithms, a training dataset was created. To get a labeled dataset of conspiracy and non-conspiracy videos, use was made of the aforementioned channel dataset made by Ledwich and Zaitsev (2020). For each channel in that dataset, the title, description, and transcript of the ten most recently uploaded videos were downloaded using YouTube's API. Videos uploaded by a conspiracy channel were then labeled as conspiracy videos, and videos uploaded by a channel from a different category were labeled as normal videos. Additionally, the channel description and channel keywords (which are used for targeted advertising on YouTube) were added to each video. The final dataset contained 65.683 unique YouTube videos, 22.156 of which were considered conspiracy videos.

Data Cleaning

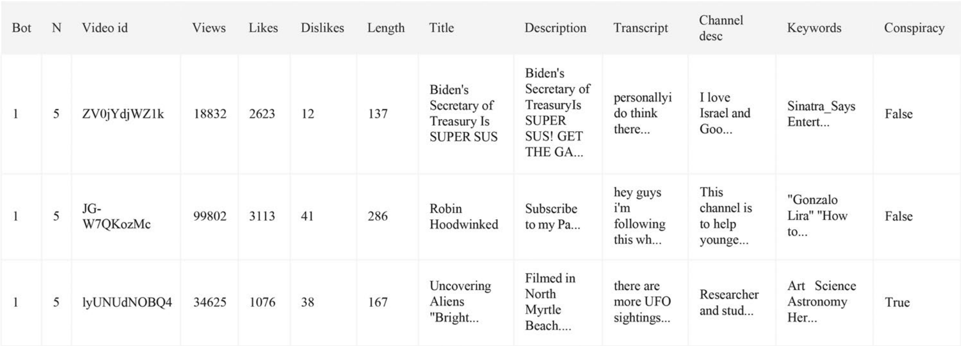
However, this dataset was not yet suitable for machine learning, as the data was still messy. Therefore, multiple steps were taken in order to clean the data. Firstly, the two classes (conspiracy and non-conspiracy) were balanced, so that the classifier would not develop a bias for non-conspiracy videos. Rather than opting for balancing the two classes through the use of class weights (a technique where weights are attributed to classes, thereby telling the classifier that getting a prediction correct for a certain, underrepresented class is more important), the choice was made to undersample the data in order to equalize both classes (both containing 22.156 videos, for a total of 44.312 videos) (Lemaître et al. 2017; Sun et al. 2006). Considering the large size of the dataset, undersampling was preferable over implementing class-weights (Brownlee 2021). After both classes had been balanced, the text for each video had to be translated into English. Since the original dataset by Ledwich and Zaitsev (2020) also contained channels by non-English speakers, these videos had to be automatically translated. Then, a few common cleaning methods were applied: all text was converted to lowercase, after which special characters, such as emojis were removed, whereafter stop words were removed and all words were stemmed using the porter stemmer (Wahiba and Karaa 2013). Finally, each video was TF-IDF vectorized to allow the classifiers to function.

Machine Learning

Table A1 shows the features used to classify videos.

The hyperparameter tuning led to impressive scores for all classifiers. When making predictions for the test set, the best-performing classifier was the support-vector machine making use of the Radial Basis Function (RBF) kernel and a penalty parameter (C-value) of 10. The SVM was tied for F1-score with the neural network using the identity activation function, with 10 hidden layers of 10 neurons. Ridge regression with a sparse-cg solver and penalty (alpha) value of 0.1 took third place, very closely followed by logistic regression with an L2 penalty, a penalty (C) value of 20, and a newton-cg solver. The worst-performing classifier was also the simplest of the bunch: the k-nearest neighbors classifier ($K = 1$). Although its performance was still formidable, it did substantially worse than the others. An overview of all metrics for each classifier can be seen in Figure 1. The ten best-performing configurations for each classifier can be found in Table A2.

Table A1. Example output of the experiment. Every row contains the metadata of a video recommended to a bot and the label from the conspiracy classifier.



Bot	N	Video id	Views	Likes	Dislikes	Length	Title	Description	Transcript	Channel desc	Keywords	Conspiracy
1	5	ZV0jYdjWZ1k	18832	2623	12	137	Biden's Secretary of Treasury Is SUPER SUS	Biden's Secretary of TreasuryIs SUPER SUS! GET THE GA...	personallyi do think there...	I love Israel and Goo...	Sinatra_Says Entert...	False
1	5	JG-W7QKozMc	99802	3113	41	286	Robin Hoodwinked	Subscribe to my Pa...	hey guys i'm following this wh...	This channel is to help younge...	"Gonzalo Lira" "How to...	False
1	5	lyUNUdNOBQ4	34625	1076	38	167	Uncovering Aliens "Bright...	Filmed in North Myrtle Beach....	there are more UFO sightings...	Researcher and stud...	Art Science Astronomy Her...	True

Table A2. Results of the classifiers on the validation set with different hyperparameters. **Ensemble.**

Ensemble	Activation	Layers	Neurons	Accuracy	Precision	Recall	F1
svm, nn, knn	logistic	1	20	**0.9393**	**0.9489**	**0.9312**	**0.9399**
svm, nn, knn	relu	1	20	0.9393	0.9489	0.9312	0.9399
svm, nn, knn	identity	1	1	0.9393	0.9489	0.9312	0.9399
svm, nn, knn	identity	10	1	0.9393	0.9489	0.9312	0.9399
svm, nn, knn	logistic	1	10	0.9393	0.9489	0.9312	0.9399
ridge, svm, nn, knn	tanh	10	1	0.9393	0.9489	0.9312	0.9399
svm, nn, knn	tanh	1	10	0.9393	0.9489	0.9312	0.9399
svm, nn, knn	tanh	1	20	0.9393	0.9489	0.9312	0.9399
svm, nn, logr, knn	identity	1	1	0.9393	0.9489	0.9312	0.9399
svm, nn, logr, knn	identity	1	10	0.9393	0.9489	0.9312	0.9399

Support-vector machine.

Kernel	C	Accuracy	Precision	Recall	F1
rbf	10.0	**0.936309**	0.945473	0.928747	**0.937035**
rbf	100.0	0.935557	0.942289	**0.930713**	0.936465
rbf	1.0	0.925276	0.930163	0.922850	0.926492
poly	10.0	0.916499	**0.946017**	0.886978	0.915547
poly	100.0	0.915246	0.944940	0.885504	0.914257
linear	1.0	0.913741	0.917944	0.912531	0.915229
poly	1.0	0.909729	0.935065	0.884521	0.909091
linear	10.0	0.904965	0.907882	0.905651	0.906765
linear	100.0	0.898195	0.905830	0.893366	0.899555
rbf	0.1	0.878887	0.878906	0.884521	0.881705

Neural network.

Activation	Layers	Neurons	Accuracy	Precision	Recall	F1
identity	10	10	**0.923019**	**0.935484**	0.912039	**0.923613**
identity	25	10	0.921013	0.933031	0.910565	0.921661
relu	10	10	0.919007	0.917561	0.924324	0.920930
identity	10	20	0.916750	0.906056	**0.933661**	0.919652
relu	10	20	0.915998	0.912221	0.924324	0.918233
tanh	10	10	0.915747	0.914592	0.920885	0.917728
relu	1	1	0.915747	0.931876	0.900737	0.916042
tanh	25	20	0.915496	0.919052	0.914988	0.917016
tanh	10	20	0.914744	0.920178	0.912039	0.916091
logistic	1	1	0.913741	0.925516	0.903686	0.914470

Ridge regression.

Solver	Alpha	Accuracy	Precision	Recall	F1
auto	0.1	**0.918506**	0.919118	**0.921376**	**0.920245**
sparse_cg	0.1	0.918506	0.919118	0.921376	0.920245
sag	0.1	0.918255	**0.919902**	0.919902	0.919902
auto	1.0	0.917252	0.923497	0.913514	0.918478
sparse_cg	1.0	0.917252	0.923497	0.913514	0.918478
sag	1.0	0.917252	0.923497	0.913514	0.918478
sag	10.0	0.878385	0.893002	0.865356	0.878962
auto	10.0	0.878134	0.892549	0.865356	0.878743
sparse_cg	10.0	0.878134	0.892549	0.865356	0.878743
auto	100.0	0.812437	0.854545	0.762162	0.805714

(Continued)

Table A2. Continued.

Logistic regression.

Penalty	C	Solver	Accuracy	Precision	Recall	F1
l2	20	newton-cg	**0.918506**	**0.924107**	**0.915479**	**0.919773**
l2	20	saga	0.918506	0.924107	0.915479	0.919773
l2	20	sag	0.918506	0.924107	0.915479	0.919773
l2	10	sag	0.916249	0.920831	0.914496	0.917653
l2	10	newton-cg	0.916249	0.920831	0.914496	0.917653
l2	10	saga	0.916249	0.920831	0.914496	0.917653
l2	10	lbfgs	0.915747	0.919506	0.914988	0.917241
2	20	lbfgs	0.914744	0.921432	0.910565	0.915966
none	1	sag	0.913992	0.923848	0.906143	0.914909
none	10	saga	0.913240	0.922461	0.906143	0.914229

K-nearest neighbors.

K	Accuracy	Precision	Recall	F1
1	**0.889669**	0.888456	0.896314	**0.892368**
3	0.888415	0.882212	0.901720	0.891859
4	0.879137	0.908899	0.848157	0.877478
5	0.873370	0.858482	0.900246	0.878868
6	0.873119	0.882441	0.866830	0.874566
2	0.872618	**0.935043**	0.806388	0.865963
7	0.868355	0.848891	0.902703	0.874970
8	0.867603	0.867382	0.874201	0.870778
9	0.861585	0.835672	**0.907125**	0.869934
10	0.859579	0.854397	0.873710	0.863946

The best score per metric is written in bold.

Noteworthy is the fact that the optimal ensemble actually outperformed the support-vector machine by a slight margin. This ensemble, consisting of the SVM, the neural network, and the k-nearest neighbor classifiers, got slightly higher scores than the runner-up across the board. The ensemble had a sixteen-way tie for best-performing parameters, all of which contained at least the SVM, neural network, and k-NN classifiers.

Though the ensemble outperformed the other classifiers, it has a significant drawback: its training time is significantly larger than that of the individual classifiers. Support-vector machines are infamous for their slowness when there is a lot of training data, and neural networks can require a lot of training time whenever the number of neurons gets large Burges and Schölkopf (1997) and Kamarthi and Pittner (1999). Requiring both algorithms to run would therefore require a lot of additional training time. Considering the marginal performance increase, the cost outweighs the benefit. As a result, when taking everything into account, the support-vector machine is the best classifier for labeling conspiracy videos on YouTube.