Is YouTube Still a Radicalizer? An Exploratory Study on Autoplay and Recommendation

Simon Markmann and Christian Grimme(✉)

Department of Information Systems, University of Münster, Leonardo-Campus 3, 48149 Münster, Germany
{s_mark11,christian.grimme}@uni-muenster.de

Abstract. This work investigates the functioning of YouTube's recommendation system with focus on the autoplay function. The autoplay function was often referred to as "radicalizer" in the past, as it was considered to lead towards more extremist content. By an automated data collection through browser remote control, we simulate different usage scenarios (allowing and disallowing autoplay) with personalized accounts as well as with anonymous users. This leads to multiple recommendation paths, which are analyzed. The presented analyses suggest that while YouTube continues to rely on familiar mechanisms for capturing users' attention, ongoing public criticism with respect to the recommendation system has seemingly led to changes in YouTube's algorithm parameterization and to more cautious recommendations.

Keywords: YouTube · Recommender system · Autoplay · Radicalization · Misinformation

1 Introduction

Recommender systems [2,7] are part of our daily use of the Internet. They are embedded in search engines, in social media, and in trading platforms. As such, they are used - usually unnoticed - by billions of users. These systems go beyond simply 'sorting' unorganized information on the Internet. Unlike early search engines from the ancient days of the Internet, they deliver individualized information (i.e., information tailored to the user or a user group). They try to deduce which information artifacts are useful and which are less helpful with respect to user preferences, semantic contexts, and behavioral patterns [1]. Their superiority in providing mostly appropriate content has largely contributed to the success and market dominance of Google as *the* search engine of our time. In fact, recommender systems are absolutely necessary components of today's platforms and often essential for them to survive in the battle for the attention of Internet users on a relevant scale.

In addition to the (attention) economic benefits of recommendation systems for platforms and users, the social problems associated with these systems are

J. Bright et al. (Eds.): MISDOOM 2021, LNCS 12887, pp. 50–65, 2021.
https://doi.org/10.1007/978-3-030-87031-7_4

increasingly being discussed and highlighted as e.g. by Stöcker [27]. These problems include the use of user interaction as a relevance signal and the misinterpretation of those signals by the recommender system. At the same time, these signals can be deliberately set from the outside to influence the recommendations of the system. The interaction of the user signals, the deliberately designed user interface, the preparation of information and, of course, commercial interests lead to a complex amalgamation that can result in misdevelopments or even radicalization. In a system that classifies user interests on the basis of user signals, captures emotions and combines them with seemingly suitable suggestions to direct attention, there can be no question of informational objectivity and freedom from bias.

YouTube, one of the world's largest video platforms, uses a recommendation system for suggesting videos [10], as do other social media platforms for suggesting other content. The stated goal of this system is, on the one hand, to offer videos to users, which match their interests or satisfy their personal need for information. On the other hand, out of economic interest, users should naturally spend as much time as possible watching videos on the (ad-supported!) platform [20]. At the beginning of 2018, YouTube's Chief Product Officer Neal Mohan stated that 70% of total video consumption (in terms of video viewing time) is due to suggestions from the recommendation system [26]. In addition to the actual video being viewed, YouTube displays other recommended videos. In 2015, YouTube also introduced an autoplay function, which automatically recommends another video at the end of a watched video and plays it automatically without user interaction [5].

However, the recommendation system of YouTube, which most of the time works inconspicuously for users, sometimes attract attention by making and realizing (in the autoplay case) recommendations that seem unusual or even frightening and dangerous. For example, in the context of the 2016 U.S. election campaign, the New York Times reported YouTube as "The great radicalizer" [31] and noted that extreme videos on YouTube quickly became part of the recommendations. The autoplay function in particular is attributed with the property of delivering radical or inappropriate content, disinformation and fake news [19,24], and in some cases even promoting a convergence toward this content [28,31] (also in the sense of a filter bubble [22]).

Also as a reaction to these reports and their public resonance, YouTube has recently announced in many blogs and articles [14,21,32] that it will react to problematic algorithmic behavior and adapt its recommender system. This should be accompanied not least by measures that promote quality content and combat fake news and disinformation.

Since YouTube's business model is of course centrally based on its recommendation system, the algorithms and possible changes are classified as a trade secret and are not disclosed. An audit of the announced measures is only carried out by YouTube itself and is difficult to perform in an independent way. From a methodological point of view, external testing of these announcements means,

above all, that the functioning of the recommender system (as a black box) must continue to be challenged on a regular basis.

This work reports a recent experimental and exploratory study focusing in particular on the autoplay mechanism and its functioning. Therefore, the study follows the autoplay recommendations for several steps "in depth" and analyzes the diversity or convergence of recommendation paths - starting from different profiles and subject areas. This is a first systematic, experimental step towards evaluating previous models, simulations, and observations as e.g. reported by Stöcker and Preuss [28].

In addition, however, other current features of the recommender system can be derived from the experimentally collected data, allowing limited insight into the recommender's operation and thus some speculation on the design and control issues that arise for YouTube with the recommender system and its public perception.

After a brief review of the literature in the context of this work in Sect. 2, the next Sect. 3 moves on to the experimental design. Section 4 presents and analyses the results. Eventually, Sect. 5 discusses the results, the necessity to further investigate recommender systems in platforms, and points to an inherent design and control problem for YouTube.

2 Related Work

Scientific analysis of commercial recommender systems faces the major problem that these systems are considered trade secrets of the companies which use them [23]. This secrecy of algorithms and processed data is essential for the economic existence of the companies whose entire business model is based on these recommender systems. In this respect, an investigation of these black box systems from the outside is always limited and only of restricted significance. At the same time, however, it is important that these investigations - be they individual observations or systematic surveys of specific aspects - are carried out. They allow small insights into complex systems such as search engines or even video platforms such as YouTube, but in their totality they can also provide a framework for simulating [28] and even evaluating these systems, including in terms of individual or societal impact [11,33].

We briefly consider here some of these approaches to what is often called auditing of recommender systems. We specifically focus on the context of YouTube, the impact of these audits on public perception, and the response of the platform itself.

2.1 Analyzing How the YouTube Recommender Might Work

As mentioned before, YouTube's recommendation system is a black box and can only be analyzed to a limited extent by external parties. An analysis always means that aspects of the system can be checked for their behavior in a very

selective way. Already very early investigations, which often focused on measuring popularity development of videos in YouTube's platform [8], relied on crawling data from the platform [8,35] or on additional measurement of network activity, e.g. at an university campus [37]. Other approaches used search queries and crawling as strategies to acquire insights into personalization [13] of content delivery by YouTube. Only very rare publications of YouTube itself allow some restricted insights into the recommender system. As such YouTube published in 2012 that it had reconfigured video recommendation to weight watch time more strongly [20]. In 2016, some developers of the recommender AI presented the basic structure of the filtering and recommendation system (using deep learning) without providing too much detail [10]. Most interesting, the recommender system is - according to the developers - parameterizable. This ensures that the YouTube product can be adjusted constantly regarding its behavior. This large dynamic of the recommender system makes a reliable analysis and an explanation of observed effects even more difficult. Therefore, on the one hand, it is important to continuously re-survey the behavior of the recommendation system [15,16]. On the other hand, approaches like those of Stöcker and Preuss [28] are worth emphasizing. Based on insights gained so far through other studies and YouTube's publications, the authors have created a simple simulation model to study the effects of autoplay and demonstrate observed effects (such as the convergence of autoplay recommendations towards problematic content).

This paper fits into the context of the ongoing investigation of YouTube's recommendation system, and at the same time tries to pick up some of the previous findings and simulation results in order to assess YouTube's development - also under the impression of the larger public perception of societal issues with recommendations.

2.2 Issues with Recommendations in YouTube

In recent years, scientific research, various experiments and newspaper reports or data journalistic investigations have had an ever-increasing impact on the perception of YouTube's recommendation system, which actually works in the background. Former employees of YouTube and journalists have analyzed the proposals of the YouTube recommender system - especially in the context of the US presidential election in 2016 - and found that this system was able to help extreme and radical content gain visibility [17,18,31]. In a paper on the major social media platforms as an ecosystem for disinformation, Stöcker [27] reports on various examples that show that the optimization of criteria such as watch time can lead to the disproportionate presentation of radical or conspiracy-theory content, as can a focus on the frequency of clicks on a video or ratings. In the context of science communication, Allgaier [3] confirms this assessment and provides another example: he experimentally demonstrates that the recommendation system disproportionately suggests video content, which contradicts the mainstream science in the context of climate change.

These insights and their media reappraisal have certainly contributed significantly to the critical view of the public on YouTube's recommendation system

(and also the systems of other platforms). In the context of YouTube, this can also be seen in two very recent surveys: in their study Zimmermann et al. [36] report on young people's consumption of YouTube content on political and social issues and find greater skepticism about the trustworthiness of the content presented there. The videos on the YouTube platform are seen as more entertaining than classically produced TV content. At the same time, however, YouTube content is also described as less objective, opinion-oriented, more emotional, and less credible. Further it is considered to be manipulative. Another study reveals that people live in an ambivalent relationship with recommendation systems [4]: On the one hand, people rarely trust the decisions of artificial decision makers. At the same time, a majority of respondents are convinced that artificially intelligent recommendation systems make better decisions than human decision-makers do.

The scientific study of YouTube's recommendation system and the media discussion partly based on it thus seem to influence also the perception of recommendation systems among users of YouTube and other platforms. It can be assumed that these reactions have contributed to the fact that, on the one hand, regulatory considerations have been made and, at the same time, containment measures have been announced by YouTube [14,21,32].

Interestingly, a recent dissertation [15] research (and conducted parallel to this study) investigates the recommender system by following suggestions in an automated way and concludes that some of the previously observed and reported anomalies (especially the convergence to critical content) are no longer present. The study could be seen as a first indication of changes in YouTube's recommendation system. However, the data basis is still very insufficient even for a preliminary statement - also due to a limited number of experiments conducted in the mentioned study. This study strives for providing further exploratory experiments and thus additional insights into the functioning of the recommendation system in order to build a first picture of YouTube's reactions and their effects.

3 Experimental Design

The recommendation system of YouTube is examined with the help of an exploratory experiment. The focus is on the video suggestions that are displayed on the right side of the website when a video is played. This investigation is intended to provide insights into the underlying algorithmic systematic of the platform. Although this work focuses on data collection along the automated recommendations of the autoplay function, additional information is collected during this process. This allows additional analysis and inductive research based on the gathered data.

Figure 1 shows the general setting of the implemented experiment. YouTube-offered videos from different categories were played with 30 personalized accounts. After finishing a video, the next suggested autoplay video was allowed to start, finally resulting in a graph of videos with start and end nodes at a

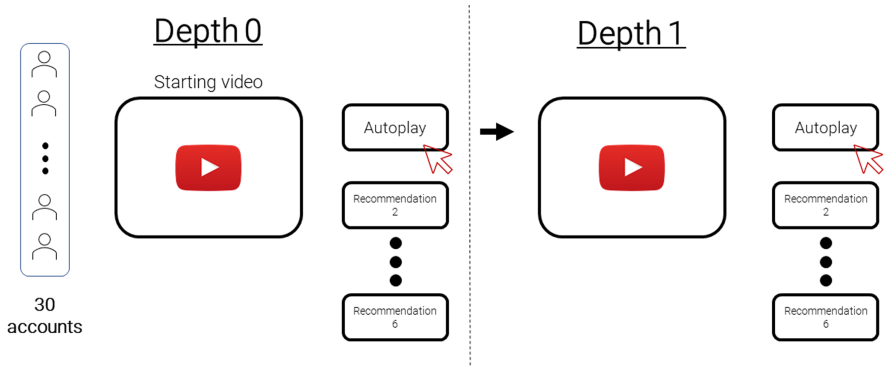

Fig. 1. The created accounts start the experiment by watching a prior chosen video in 'depth 0' and play the first recommended video afterwards, which is the video of the autoplay function. This process is executed for each run until 'depth 10'.

depth of 10. The watched videos were saved along with the top-five video recommendations at each stage. All relevant information about the collected videos, the users, and the runs were stored in a database to be analyzed afterwards.

This setting provided a database of almost 30,000 video recommendations collected between February and March 2021 to expose underlying mechanisms of YouTube's recommender system. These videos were automatically played on a server of the University of Münster, partly with logged-in accounts and partly without accounts. As the server was operated behind a shared IP address, watch histories cannot be allocated directly to one person by YouTube. At the same time, however, this limits the variety of possible users to people who have access to this network, to academics and students, making it not representative for the German population.

Out of these 30 accounts 28 were manually created while two had already been actively used before. The accounts were created with the intention to appear as realistic as possible and also to represent a sample size of randomly chosen YouTube users. To achieve this, first personal profiles were created for each account, which took into consideration demographic aspects like age or gender as well as personal interests. The average age of the users was 35 with more young users than older ones and the genders were evenly distributed. These aspects were aligned to a statistic of YouTube user demographics [29].

Subsequently each of the accounts watched ten to fifteen videos in a clean browser to give the recommender system a chance for classifying them according to their account properties and interests. 'Clean' means that the browser data got deleted every time an account watched a couple of videos and the users changed web browsers in order to hinder YouTube in spotting links between accounts. Each account played videos that were partly random and partly inspired by their demographic aspects. Here the goal was to get many distinct video topics but also some overlapping ones. In the end each user had his or her own distinct

starting page with video recommendations of which the first 20 videos also got collected.

The users (or no user) played each video for a relative amount of time. This amount was either 5%, 50% or 95% of the videos duration. After the completion of each run, the search history got deleted which led to no traces of activity in the recommendations of the starting pages.

To determine whether the recommender system behaves different for individual video categories, four starting videos were chosen. The first video (News) was selected as a daily news video from the German channel ZDF, which is a public broadcaster. The next video (Music) was a music video of a song by the American artist Post Malone. The third video (Covid) dealt with the Covid-19 crisis. The video cannot be found directly via the search bar anymore and it criticizes a famous German virologist and the WHO. The last video (Trend) was a short video from the trend section that got re-uploaded from the platform TikTok. It shows a family doing a funny challenge. Time and capacity limits lead to the conclusion that four starting videos would be the maximum. The categories 'News' and 'Covid' are of social and political interest and center of critique which is why they were chosen, also because 'News' is a big section on YouTube. 'Music' is the category with YouTube's most watched videos. 'Trend' is full of videos that are currently famous and full of different types of created content.

For the automated data collection, a script was created, which initiated a Google Chrome web driver with the help of the test-software Selenium. This driver can start either with or without the user data of the accounts. This also allows video suggestions to be collected that are not dependent on the users' profiles. The script can then be executed for different starting videos, numbers of runs and percentages of watch time and it collects the URL addresses of the autoplay videos and five other recommended videos. The autoplay video is the video, which is recommended first for each watched video and starts automatically after a video ends, if the autoplay function is not deactivated. For the automated skipping of advertisements, an additional browser extension was downloaded from the Chrome Web Store, customized with the help of another script, and added to the web driver. The video URLs were then fed into the YouTube Data API and the outputs were saved with the user data into a SQL-based database.

4 Experiments

The following chapter provides insights into the collected data. Mainly, the diversity of the recommendations was compared to each other, the starting page, the relative playing time, the autoplay function or to the non-personalised suggestions. Furthermore, metadata such as video length, ratings or number of views were examined.

What Are the Effects on the Length of the Videos?
In the following, the influence of the various parameters on the duration of the suggested videos is analyzed. For this part of the evaluation, all videos longer than two hours were omitted, as it is unrealistic that overlong videos will be watched in their entirety. These long videos are often summarized live-streams or music compilations. In addition, these outliers strongly distort the statistics of the data.

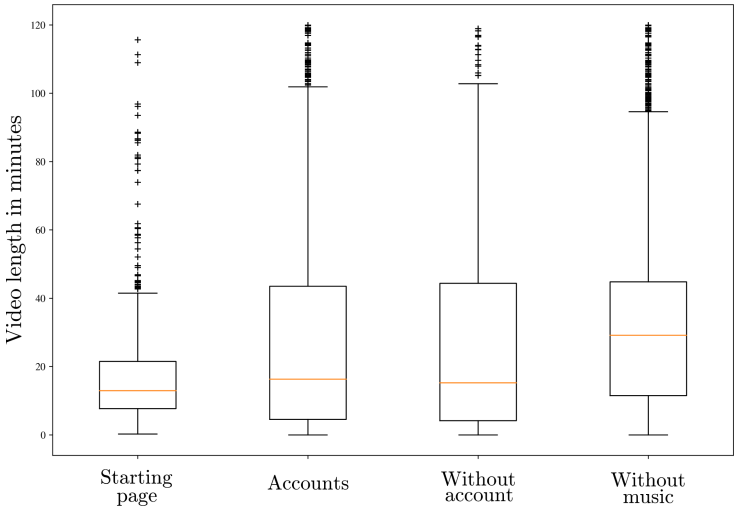

Fig. 2. The video lengths of the proposals seem to be generally longer than the videos of the starting pages. Especially without the category 'Music' there is a sharp difference.

Figure 2 shows box plots of the lengths of the videos of the users' starting pages, the recommended videos of the users, in comparison those without logged-in users and the recommendations without the category 'Music'.

We will exclude the category 'Music' oftentimes during the following evaluations, since the system acts significantly different for this category. About 73% of the channels that appear in this category have the channel tag 'Music' and most of the videos are between three and five minutes long. This segmentation makes sense for both YouTube and the user, since YouTube is used as a music platform by a lot of users and music videos have the most views [30].

It is apparent from Fig. 2 that the suggested videos are longer compared to those on the starting page, especially without the category 'Music'. Note that the videos of the starting pages (without the omitted overlong videos) are 18:30 min long at median *and* the starting videos of the three categories are also each shorter than 16 min. There is no significant difference between the personalized and non-personalised suggestions.

Additionally the video lengths of the runs in which videos were played for a short, medium, or long amount of time are compared. The category 'Music'

is again omitted because the data of this category scatters minimally around three to five minutes and thus weakens the effects for the other categories. The difference in the mean values of the long and short watched videos is 9:59 min. The Wilcoxon rank sum test for the median video lengths of the runs as random variables yields a p-value of 4.5277×10^{-117}. The hypothesis that the median lengths of the long and short watched video runs is the same, can therefore be rejected.

How Diverse Are the Recommendations for Different Users?
Table 1 shows for each depth of the runs how many of the recommended videos were suggested to more than one of the 30 users. The measured values were further subdivided into the four categories and the relative amount of video watching. The values range from 0 to 100%, where 100 (%) states that every video recommendation appeared in all of the thirty users. A value of 0 (%) on the other hand indicates that no video recommendation appeared more than once. Intermediate values should be interpreted accordingly. If there is an effect that groups the 30 users and could thus indicate filter bubbles or similar undesirable effects, then the percentage values should at least not flatten out completely. However, as can be seen, especially in the 'trend' category, the common video suggestions seem to disappear completely.

Table 1. Proportion (in %) of **videos** occurring multiple times per depth of a run across the 30 subjects.

Depth of run	News		Music		Trend		Covid-19	
	Short	Long	Short	Long	Short	Long	Short	Long
1	43	47	75	74	24	26	22	36
2	14	22	53	67	23	21	16	19
3	13	14	38	58	16	18	12	18
4	13	11	33	51	9	12	9	17
5	10	14	21	44	6	10	13	12
6	14	15	10	43	3	4	11	13
7	7	12	9	40	6	1	9	11
8	7	14	7	37	3	2	6	12
9	8	13	10	31	1	0	6	11
10	7	11	7	26	2	1	1	11

The partly high percentages for the music videos can again be explained by the fact that YouTube seems to be transforming into a music platform for these videos. In this case, similar songs are played rather than suggesting user-specific videos. For news videos, there seem to be cases of groupings. These connections rarely involve more than two users and are rather caused by the

fact that there is only a limited amount of channels that get recommended in this category. In general, for news videos, it is almost mainly documentaries or news from other news channels that are suggested. As long as videos are watched for a long time, videos from a similar category continue to be suggested. In the category 'Covid-19', too, mainly documentaries or talk shows were suggested, which rarely contained inappropriate or false content. The only user for whom there were partially critical recommendations, as evidenced by the fact that the suggestions were politically one-sided, is the one who also played exclusively one-sided political content in the course of the personalization. A comparison of the channels that uploaded the video recommendations in the two similar-looking categories returns a rather high value of 0.652 for the cosine similarity[1]. In comparison, the categories 'News' and 'Trend' have a cosine similarity of 0.043.

Table 2. Proportion (in %) of **channels** occurring multiple times per depth of a run across the 30 subjects.

Depth of run	News		Music		Trend		Covid-19	
	Short	Long	Short	Long	Short	Long	Short	Long
1	74	79	73	69	25	32	49	54
2	62	62	57	69	21	23	42	46
3	50	53	46	63	19	18	46	54
4	45	54	46	56	20	16	40	47
5	40	51	31	51	13	13	35	38
6	44	49	20	53	13	10	34	39
7	43	51	20	52	11	3	33	40
8	35	48	15	56	6	6	27	32
9	31	54	20	53	5	0	27	30
10	31	52	14	45	5	3	17	35

The same methodology was used for the channels that uploaded these videos. Table 2 shows how many of the channels were recommended more than once per depth. Again a value of 100 (%) would indicate that the same channels got recommended for all of the 30 accounts per depth of the run and a value of 0 (%) would show no same channel recommendations. It was expected that the values would be at least as high as those of the videos. In a few places this is not the case, as some videos were deleted by the platform before they were inserted into the API and therefore do not appear for the channels, but for

[1] The cosine similarity [25] is computed as angle between two vectors, which represent the frequency distribution of video tags in the compared categories. A value of 0 denotes maximum dissimilarity, while a value of 1 denotes equality.

the videos. Once again, there is a slightly recognizable difference between the different relative duration of the played videos.

Does It Matter How Long Videos are Watched?
Another investigation was conducted on the consumption duration of watched videos, which collected the average number of unique channels per total run. This metric is meaningful because a low number of unique channels cannot in any way indicate high diversity. The long and short watched videos respectively provide a mean of 20.02 and 28.97 unique channels per run. This means that the 60 videos collected in one run originate on average from 20 different channels in the first case and from 29 different channels in the second case. The Wilcoxon test returns a p-value of 8.5433×10^{-20} for the two distributions. The hypothesis that long or short viewing has no influence on the number of unique channels per run can be rejected at any relevant level of significance. The recommendation system notices when a user skips videos, considers this as negative feedback, and tries (as a kind of compensation strategy) to vary the content, or at least the channels for matching the users interests (again).

Does the Autoplay Function Act Different?
The autoplay function was analyzed by comparing the channel that uploaded the next video with the previous one. If video A from channel C starts and video B from channel C is suggested next, we assume that the suggestions do not differ strongly. For the autoplay videos, this happens 57.17% of the time for the 30 accounts. Even without users, this value is in a similar range at 58.75%. For the remaining five video suggestions of each run, the values are 29.62% and 32.23% respectively. The Wilcoxon rank sum test gives a p-value of 4.2642×10^{-45} for accounts and 1.6373×10^{-16} without an account. The random variables in this test are the values of the average consecutive same channels in the runs. This shows that the autoplay function tends to suggest more videos of the same channel. Breaking this further down into categories shows that this effect is slightly stronger for news videos (61.11%) and weaker for music videos (50%). There is also a difference for the breakdown between long and short viewing: 64.5% of the channels match their previous one for long viewing, whereas it is only 49.58% for short viewing. Testing the hypothesis that short and long viewing have an equal median for this random variable can be rejected with a p-value of 7.4756×10^{-9}. The measured values of the autoplay condition are generally high and accordingly also had an influence on the other results, since in each case the autoplay video was played next. If one of the other five videos had been played instead, the collected data and the resulting analysis could have deviated strongly.

How Does the System Change for a Personalized Account?
The difference between video recommendations with or without an account in terms of content diversity has not been considered, yet. Once again, the average number of unique channels per run was calculated. This is $\mu = 24.08$ with a standard deviation of $\sigma = 8.97$ for accounts and $\mu = 23.11$ with $\sigma = 8.09$ for

no account. Under these circumstances, the use of an account does not seem to have a great influence on the calculated parameter. A comparison of the collected channels of the video suggestions with or without an account results in a value of 0.748 for the cosine similarity. This high value can be justified with the already mentioned characteristics of the categories 'Music', 'Covid-19' and 'News'. Moreover, the runs started on the same videos.

Ratings, Clicks and Content Partnerships
The ratio of likes to dislikes for the video suggestions is 4:1, that is, 79% of the ratings are positive. Note that only 1.86% of the videos were rated. For 85% of the video suggestions there are even more than 90% positive ratings and only 0.5% of the suggestions have more negative than positive ratings. The latter are mainly videos about Germany's Covid-19 policy.

The average number of views is 138,718,992. This value breaks down for the four categories as follows: 'Music': 527,544,424, 'News': 1,424,238, 'Covid-19': 2,085,894 and 'Trend': 23,389,197. For each of the categories, the value is increased compared to the initially chosen video. The videos of the users' home pages were viewed on average 20,119,423. Moreover, the platform seems to favor the biggest channels when it comes to the order of suggestions and directs the user to videos from these channels via autoplay. For example, for the category 'Trend', the three channels that account for the most suggestions in depth 10 have an average subscriber base of 32 million, placing them among the largest channels on YouTube; 75.32% of the suggested videos are from channels that have a partnership with YouTube that allows them to monetize their videos.

5 Discussion and Conclusion

The results of this study show that YouTube has retained features of the recommendation system in many areas despite criticism from outside. As already confirmed in previous studies and also by YouTube itself, the recommendation system tends to suggest longer videos starting from the initial video. At the same time, the consumption duration of videos is used as a rating or satisfaction measure as feedback to the system. This is also important because the percentage of direct user ratings per video view is only about 1.86%. Thus, individual rating responses by users would not be sufficient as feedback for the recommender.

At the same time, however, the study showed that playing videos with and without an account has neither a significant influence on the diversity of the suggestions nor on the length of the videos. However, due to the fact that the recommendation system includes the watching history of an account, the suggestions for users with logged in account fit well with the previously set interests. It can be speculated that the basic functionality of suggestion generation is the same, but history is an important influencing factor - also in order to fit content to user preferences and to preserve their attention and interest in watching further content.

When looking at the autoplay function, we were able to show that this leads to a lower diversity of suggestions. However, the restriction of diversity refers to the channels, not videos. Again, the selection is influenced by the overall account watching history, but the immediate history also seems to be weighted more heavily.

It is also noticeable that the developers of the recommendation algorithm apparently provide quite different parameterizations for different categories. While YouTube apparently 'mutates' into a music platform for the music category, news channels and channels critical of governmental Covid-19 measures, for example, are treated very similarly when it comes to suggesting content. Especially in the last category, mainstream content (news, talk shows, and documentaries) is suggested to compensate for criticism presented. This can be interpreted as a manifestation of YouTube's responses to persistent criticism of the recommendation system.

The present results (cautiously) suggest that YouTube's interventions in the recommendation system show some effects. By its own admission, YouTube is very active in combating problematic content[2]. Its latest transparency report shows that YouTube has already removed more than 2 million channels and close to 10 million videos in the first three months of 2021[3]. Both YouTube's response and consideration of the characteristics of the recommendation system allow for two conclusions:

1. The continuous external analysis and monitoring of the Black Box recommendation system, despite their mostly exemplary nature, can be helpful to uncover problems and needs for action and thus influence (also through public discussion) the development of these systems. This can and should motivate scientists and journalists to continue to critically question the decisions of automated systems.
2. The investigation and the apparent adaptation of the decision-making system, while at the same time maintaining various optimization goals of the recommender system, suggest that platforms like YouTube are in a dichotomy between public pressure due to decision errors and the economic necessity of the recommender system (and thus the necessity to accept deviations from 'normal'). If YouTube follows the rules of exploiting users' attention (see also attention economy [6,12,34]), it cannot completely dispense with stimuli that keep users completely away from surprising new content - perhaps even content tending toward the extreme or sensational - in an attention-binding spiral [11]. Therefore, YouTube (unless regulations force it) will certainly continue to offer problematic content. Filtering content only when it enters the system or when moderation is triggered by users are only partly effective measures.

[2] https://blog.youtube/news-and-events/more-information-faster-removals-more/.
[3] https://transparencyreport.google.com/youtube-policy/removals?hl=en.

Limitations
Just like any other study of black box systems, the presented experiments have limitations. The collection of recommendations originated from four different starting videos. For other videos the results could be different. Furthermore, the results can only be verified for the time span of February till March of 2021 as afterwards the system could have changed significantly (due to new parameterization by YouTube). Also the sample size is rather small and could be influenced subjectively as the accounts were created manually. The assumption, that users always follow the autoplay function is also possibly unrealistic. Although it was the explicit focus of this study, several (possibly more realistic) configurations of user interaction are still open to be investigated.

Future Work
The data collection methodology presented here (and related methods in previous work, e.g. [15]) provides the ability to automatically and systematically collect data on the behavior of YouTube's recommender system. It is important that analogous experiments are repeated to verify the presented analyses and to track developments in the YouTube system over time. At the same time, however, analyzing the effects and the impact of scientific and media reports on how recommender systems work in an interesting field of research beyond statistical evaluation. The question of how people deal with machine decisions and which social and concrete economic effects these systems have needs to be further investigated from a diverse and interdisciplinary perspective (see e.g. [9] on open aspects in algorithmization, attention economy, and ethics).

Acknowlegments. Both authors appreciate the support of the European Research Center for Information Systems (ERCIS).

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005). https://doi.org/10.1109/TKDE.2005.99
2. Aggarwal, C.C.: Recommender Systems: The Textbook. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29659-3
3. Allgaier, J.: Science and environmental communication on YouTube: strategically distorted communications in online videos on climate change and climate engineering. Front. Commun. **4**, 36 (2019). https://doi.org/10.3389/fcomm.2019.00036
4. Araujo, T., Helberger, N., Kruikemeier, S., de Vreese, C.H.: In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc. **35**(3), 611–623 (2020). https://doi.org/10.1007/s00146-019-00931-w
5. Brinkmann, M.: Google tests new video autoplay feature on YouTube (2015). https://www.ghacks.net/2015/01/28/google-tests-new-video-autoplay-feature-on-youtube/. Accessed 29 Mar 2021
6. Brynjolfsson, E., Oh, J.: The attention economy: measuring the value of free digital services on the internet. In: ICIS 2012 Proceedings (2012)

7. Burke, R., Felfernig, A., Göker, M.H.: Recommender systems: an overview. AI Mag. **32**(3), 13–18 (2011). https://doi.org/10.1609/aimag.v32i3.2361
8. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC 2007, New York, NY, USA, pp. 1–14. Association for Computing Machinery (2007). https://doi.org/10.1145/1298306.1298309
9. Coombs, C., et al.: What is it about humanity that we can't give away to intelligent machines? A European perspective. Int J. Inf. Manag. **58** (2021). https://doi.org/10.1016/j.ijinfomgt.2021.102311
10. Covington, P., Adams, J., Sargin, E.: Deep neural networks for YouTube recommendations. In: Proceedings of the 10th ACM Conference on Recommender Systems, RecSys 2016, New York, NY, USA, pp. 191–198. Association for Computing Machinery (2016). https://doi.org/10.1145/2959100.2959190
11. Eyal, N., Hoover, R.: Hooked - How to Build Habit-Forming Products. Penguin Publishing Group, New York (2014)
12. Goldhaber, M.H.: The attention economy and the Net. First Monday (1997). https://doi.org/10.5210/fm.v2i4.519
13. Hannak, A., et al.: Measuring personalization of web search. In: Proceedings of the 22nd International Conference on World Wide Web, WWW 2013, New York, NY, USA, pp. 527–538. Association for Computing Machinery (2013). https://doi.org/10.1145/2488388.2488435
14. Hern, A.: YouTube to manually review popular videos before placing ads, January 2018. http://www.theguardian.com/technology/2018/jan/17/youtube-google-manually-review-top-videos-before-placing-ads-scandal-logan-paul
15. Heuer, H.: Users & machine learning-based curation systems. Ph.D. thesis, University of Bremen, Bremen, July 2020
16. Hussein, E., Juneja, P., Mitra, T.: Measuring misinformation in video search platforms: an audit study on YouTube. Proc. ACM Hum. Comput. Interact. **4**(CSCW1), 1–27 (2020)
17. Lewis, P.: 'Fiction is outperforming reality': how YouTube's algorithm distorts truth, February 2018. http://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth
18. Lewis, P., McCormick, E.: How an ex-YouTube insider investigated its secret algorithm, February 2018. http://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot
19. Maack, M.: 'YouTube recommendations are toxic', says dev who worked on the algorithm (2019). https://thenextweb.com/google/2019/06/14/youtube-recommendations-toxic-algorithm-google-ai/. Accessed 26 Mar 2021
20. Meyerson, E.: YouTube now: why we focus on watch time (2012). https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/
21. Newton, C.: YouTube says it will recommend fewer videos about conspiracy theories, January 2019. https://www.theverge.com/2019/1/25/18197301/youtube-algorithm-conspiracy-theories-misinformation
22. Pariser, E.: The Filter Bubble: What the Internet is Hiding From You. Penguin, London (2011)
23. Pasquale, F.: The Black Box Society. Harvard University Press, Cambridge (2015)
24. Rieder, B., Matamoros-Fernández, A., Coromina, Ò.: From ranking algorithms to 'ranking cultures' investigating the modulation of visibility in YouTube search results. Convergence **24**(1), 50–68 (2018)

25. Singhal, A.: Modern information retrieval: a brief overview. IEEE Data Eng. Bull. **24**(4), 35–43 (2001)
26. Solsman, J.E.: YouTube's AI is the puppet master over most of what you watch (2018). https://www.cnet.com/news/youtube-ces-2018-neal-mohan/. Accessed 8 Mar 2021
27. Stöcker, C.: How facebook and google accidentally created a perfect ecosystem for targeted disinformation. In: Grimme, C., Preuss, M., Takes, F.W., Waldherr, A. (eds.) MISDOOM 2019. LNCS, vol. 12021, pp. 129–149. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39627-5_11
28. Stöcker, C., Preuss, M.: Riding the wave of misclassification: how we end up with extreme YouTube content. In: Meiselwitz, G. (ed.) HCII 2020. LNCS, vol. 12194, pp. 359–375. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49570-1_25
29. Tankovska, H.: Most popular social networks worldwide as of January 2021, ranked by number of active users (2021). https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Accessed 24 Mar 2021
30. Tankovska, H.: Most popular YouTube videos based on total global views as of February 2021 (2021). https://www.statista.com/statistics/249396/top-youtube-videos-views/. Accessed 30 Mar 2021
31. Tufekci, Z.: YouTube, the Great Radicalizer. The New York Times, March 2018. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html
32. Waterson, J.: YouTube bans videos promoting Nazi ideology, June 2019. http://www.theguardian.com/technology/2019/jun/05/youtube-bans-videos-promoting-nazi-ideology
33. Williams, J.: Stand Out of Our Light: Freedom and Resistance in the Attention Economy. Cambridge University Press, Cambridge (2018). https://doi.org/10.1017/9781108453004
34. Zhang, Y., Goh, K.H.: Attracting versus sustaining attention in the information economy. In: Cho, W., Fan, M., Shaw, M.J., Yoo, B., Zhang, H. (eds.) WEB 2017. LNBIP, vol. 328, pp. 1–14. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99936-4_1
35. Zhou, R., Khemmarat, S., Gao, L.: The impact of YouTube recommendation system on video views. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, IMC 2010, New York, NY, USA, pp. 404–410. Association for Computing Machinery (2010). https://doi.org/10.1145/1879141.1879193
36. Zimmermann, D., et al.: Influencers on YouTube: a quantitative study on young people's use and perception of videos about political and societal topics. Curr. Psychol. (3), 1–17 (2020). https://doi.org/10.1007/s12144-020-01164-7
37. Zink, M., Suh, K., Gu, Y., Kurose, J.: Characteristics of YouTube network traffic at a campus network - measurements, models, and implications. Comput. Netw. Int. J. Comput. Telecommun. Netw. **53**(4), 501–514 (2009). https://doi.org/10.1016/j.comnet.2008.09.022