Technologically scaffolded atypical cognition:

The case of YouTube's recommender system

Mark Alfano, Macquarie University & Delft University of Technology

Amir Ebrahimi Fard, Delft University of Technology

J. Adam Carter, University of Glasgow

Peter Clutton, Australian National University

Colin Klein, Australian National University[1]

[1] Authorship note: M.A. designed the study, coded data, and wrote the majority of the manuscript. A.E.F. wrote the code to collect data from YouTube and made the figures. J.A.C. coded data and wrote and edited much of the manuscript (especially the discussion). P.C. coded data and provided literature review. C.K. ran the topic models and wrote and edited much of the manuscript. Work partly supported by Australian Research Council grant DP190101507 (to M.A. and C.K.) and Templeton Foundation grant 61387 (to M.A.).

Abstract: YouTube has been implicated in the transformation of users into extremists and conspiracy theorists. The alleged mechanism for this radicalizing process is YouTube's recommender system, which is optimized to amplify and promote clips that users are likely to watch through to the end. YouTube optimizes for watch-through for economic reasons: people who watch a video through to the end are likely to then watch the *next* recommended video as well, which means that more advertisements can be served to them. This is a seemingly innocuous design choice, but it has a troubling side-effect. Critics of YouTube have alleged that the recommender system tends to recommend extremist content and conspiracy theories, as such videos are especially likely to capture and keep users' attention. To date, the problem of radicalization via the YouTube recommender system has been a matter of speculation. The current study represents the first systematic, pre-registered attempt to establish whether and to what extent the recommender system tends to promote such content. We begin by contextualizing our study in the framework of *technological seduction*. Next, we explain our methodology. After that, we present our results, which are consistent with the radicalization hypothesis. Finally, we discuss our findings, as well as directions for future research and recommendations for users, industry, and policy-makers.

Keywords: technological seduction, transformative experience, radicalization, YouTube, recommender systems, conspiracy theory

Word count: 9291

Technologically scaffolded atypical cognition:
The case of YouTube's recommender system

Introduction

On January 6, 2019, Buckey Wolfe killed his brother by stabbing him in the head with a four-foot long sword. He then called the local police to turn himself in, saying that he had acted because he thought his brother was a lizard.[2] This appears to have been an expression of his belief in the so-called reptilian conspiracy theory, first popularized by David Icke (1999), which holds that shape-shifting alien reptiles live among us and control world events. Reptilianism was not the only extremist conspiracy theory that Wolfe seems to have accepted; he was also deeply enmeshed in the QAnon conspiracy theory and was an avid supporter of the far-right Proud Boys

[2] For news coverage see https://www.seattletimes.com/seattle-news/crime/god-told-me-he-was-a-lizard-seattle-man-accused-of-killing-his-brother-with-a-sword/ and https://www.huffpost.com/entry/proud-boy-allegedly-murders-brother-with-a-sword-thinking-hes-a-lizard_n_5c36042ee4b05b16bcfcb3d5.

organization.[3] A short investigation by a citizen journalist followed the pattern of videos that Wolfe "liked" on the YouTube streaming video platform, revealing that over the course of a few years, his interests shifted from music to martial arts, fitness, media criticism, firearms and other weapons, and video games.[4] From there, Wolfe seems to have gotten hooked on alt-lite and alt-right political content, and then eventually a range of conspiracy theories.

We take it as uncontroversial that the ideation associated with these bizarre conspiracy theories constitutes atypical cognition[5], in the sense that it is epistemically counter-normative.[6] In this paper, we address the question to what extent *technological scaffolding* may be to blame for such ideation, at least in some cases. While the example of Wolfe is dramatic and we do not wish

[3] See https://www.huffpost.com/entry/proud-boy-allegedly-murders-brother-with-a-sword-thinking-hes-a-lizard_n_5c36042ee4b05b16bcfcb3d5.

[4] See https://threadreaderapp.com/thread/1083437810634248193.html.

[5] In fact, as it's been argued in recent work on conspiracy theories by Cassam (2019), the kind of 'built-in' implausibility of paradigmatic conspiracy theories makes them such that belief in them will — generally speaking — require a subversion of rational norms. See also Lewandowsky, Kozyreva, & Ladyman (2020).

[6] Though see Levy (2019) for an argument that, from the inside, conspiracy theorizing is not irrational. Even if this is true in general, Wolfe's case clearly represents some sort of normative failing. And even when conspiracy theorizing is subjectively reasonable, there is often something objectively irrational about it. For discussion of this latter point, see Simion et al. (2016), among others.

to make an individual diagnosis, it seems to fit into a broader pattern of what Alfano, Carter, & Cheong (2018) call technological seduction. In recent years, academic critics such as Zeynep Tufekci and technological whistleblowers such as Guillaume Chaslot have raised the alarm about technological scaffolds that have the potential to radicalize the people who interact with them.[7] Anecdotal reports of YouTube recommending fake news, conspiracy theories, child pornography, and far right political content have cropped up in North America, South America, Europe, and elsewhere.[8] Ribeiro et al. (2018) examined alt-right recommendations on YouTube and found that there is a pathway from other types of content to these topics. However, to date, these concerns have been speculative and anecdotal, as there have been no systematic studies of the promotion and amplification of conspiracy theories via the YouTube recommender system. In this paper, we fill that gap with a large-scale, pre-registered exploration of the YouTube recommender system.[9] We aim to establish a how-possibly explanation of radicalization through the YouTube recommender system; that is, we aim to show that there exists a robust pathway from certain seemingly anodyne topics to conspiracy theories via the recommender system.[10]

[7] See https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html and https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth.

[8] See https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html.

[9] The pre-registration and other details can be found at https://osf.io/cjp96/?view_only=e56fc77588194336810d41aeef04f822.

[10] For more on how-possibly explanations, see Resnik (1991). For another recent attempt to offer how-possibly explanations of troubling social phenomena, see O’Connor (2019).

Whether this pathway is one that actual users follow (and what proportion of them do so) is left to future research.[11]

Here is the plan for this paper: we begin in section 1 by explaining the theory of technological seduction in terms of technologically-mediated cognitive pressures and nudges that subtly but systematically induce acceptance of problematic beliefs. Using this framework, we establish a hypothesis to be tested. In particular, we examine the extent to which watching videos about the following topics is liable to lead to recommendations of conspiracy theories and other problematic content: martial arts, fitness, firearms, natural foods, tiny houses, and gurus such as Jordan Peterson and Ben Shapiro. We predict that some topics (especially gurus) are much more likely to lead to conspiracy theories than others. Next, in section 2 we explain our methodology, which employs a modified version of a tool that the whistleblower Chaslot built to explore YouTube's recommender system. In section 3, we test the hypothesis articulated in section 1. We find partial support for this hypothesis, which suggests that some types of content are especially liable to send viewers down a conspiratorial rabbit hole. We conclude by discussing the implications of this research, as well as directions for future research.

[11] We should note that we do not mean to treat Buckey Wolfe as a representative sample of all YouTube viewers. We doubt that all viewers are equally susceptible to self-radicalization online. Rather, it seems likely that people who are already psychologically vulnerable would be most liable to radicalization and accepting conspiracy theories. It could be that such vulnerable individuals would end up radicalized even without interacting with the YouTube recommender system; we have no way of addressing the plausibility of that counterfactual. Thanks to an anonymous referee for raising this point.

1 Technological seduction

According to YouTube, in 2018 approximately 70% of all watch-time spent on the site was driven by the recommender system.[12] If the recommender algorithm is leading people towards conspiracy theories, that would be troubling. It would also be an instance of a more general process that Alfano, Carter, and Cheong (2018) have dubbed *technological seduction.*

Alfano, Carter, & Cheong (2018) follow Forrester (1990) in thinking of seduction as a transformative experience (Paul 2014) that includes an important epistemic component. According to Forrester (1990, p. 42), "The first step in a seductive maneuver could be summed up as, 'I know what you're thinking'." Alfano and colleagues point out that recommender systems simulate such an assertion. For instance, when Google uses predictive text to suggest search queries to users, it effectively tells them, "We know what question you want to ask." Furthermore, after a search is run, Google's prescriptive results suggest the most relevant or "correct" answers to the query just run, effectively telling users, "And this is the answer to your question." In a successful seduction, the seducee accepts the assertion that his or her thoughts are known, and may end up following the seducer further both epistemically and practically.

Alfano and colleagues distinguish two types of technological seduction: *top-down* and *bottom-up*. In this paper, we are concerned only with bottom-up seduction, which occurs as the result of technological systems creating suggestions based on aggregated user data. The aggregation can be done both by combining across different users to find common patterns and

[12] See https://www.cnet.com/news/youtube-ces-2018-neal-mohan/.

by personalizing for each user based on their location, search history, and other data (what is sometimes referred to as their “digital footprint”).[13] In so doing, bottom-up seduction takes a user’s own record of engagement as the basis for saying, “I know what you’re thinking.”[14] Moreover, the YouTube recommender system is optimized primarily for watch-through, meaning that the system aims to recommend videos that users are likely to watch all the way to the end.[15] The reason for this is simple: YouTube makes money by selling advertisements displayed adjacent to or embedded in videos. To maximize ad revenue, YouTube needs to ensure that users spend as much time as possible watching and then letting the interface automatically play the next recommended video (AutoPlay is the default setting).

Let’s now consider why this might be a recipe for unintended bad consequences. As far back as Zajonc (1968), psychologists have investigated the *mere familiarity effect*. This term refers to the fact that people tend to develop positive associations with the things, people, and concepts to which they’ve been directly exposed. In Zajonc’s work, the positive association was

[13] In contrast with bottom-up technological seduction, top-down technological seduction occurs via the manipulative framing of online choice architecture that structures the user’s perception of the relevant option space in a way that guides them in certain prescribed directions. See, along with Alfano et al. (2018), also Weinmann et al. (2016) for discussion of digital nudging.

[14] See King (2019) for a similar account that is framed in terms of the presumptuousness of recommender systems.

[15] See https://www.businessinsider.com/youtube-watch-time-vs-views-2015-7?international=true&r=US&IR=T.

affective: people tended to *like* or *enjoy* things that they'd encountered before. And, as Johansson et al. (2005), Hall, Johansson & Strandberg (2012), and Hall et al. (2013) have more recently shown, the seductive framing of exposure (i.e., telling someone what mental state they embody) is an effective way to transform people's preferences and even their voting intentions. Beyond the affective or preferential manifestation of mere familiarity, there is also a doxastic or epistemic manifestation: people tend to *believe* or *think they know* the things that they've encountered before (Prentice, Gerrig, & Bailis 1997; Wheeler, Green, & Brock 1999; Ecker et al. 2011). In a suitably-constructed epistemic environment, this is an efficient and effective heuristic (Gigerenzer & Goldstein 1996, Gigerenzer 2008), but it can also go haywire in a hostile epistemic environment (Alfano & Skorburg 2018).

These and related considerations lead Neil Levy (2017) to caution against consuming fake news and other sources of falsehoods: even if one approaches them with caution, one is liable to be taken in. Personalization algorithms like the YouTube recommender system can further amplify this effect by suggesting content that is not only likely to please, but that is maximally similar to content that the user already finds engaging. If this is on the right track, then a burning question arises: are there *prima facie* unproblematic starting points which reliably lead, via the recommender system, to problematic content? In the case of Buckey Wolfe, it appears that he traversed from music to toxic masculinity and far-right politics, and thence to conspiracy theorizing. That suggests one potential pathway from the political right, from martial arts, fitness, firearms, and gurus to conspiracy theories. All of these topics are stereotypically masculine (all of the gurus are men, and many of them also have disproportionately male followings) and right-wing. To round things out, we wanted to include additional topics that are not stereotypically masculine and right-wing, but are still potentially on the fringe. After some discussion among the authors, we settled on two additional potential pathways to conspiracy theories: natural foods and tiny houses, which are associated with anti-capitalism and concerns about environmental and climate impact. It has already been established that interest in certain topics is more likely than others to lead to conspiracy theorizing on the Reddit platform (Klein, Clutton, & Dunn 2019). Could the same be true on YouTube?

After some internal discussion and a few unsystematic searches of YouTube, we arrived at some hunches about where the conspiracy theories were most likely to arise. Our pre-registered prediction is that topics are associated with the recommendation of conspiracy theories, in descending order, as follows: gurus = firearms > natural foods > martial arts > fitness > tiny homes. In other words, we predict that gurus and firearms will be most associated with conspiracy theory recommendations, followed by natural foods, martial arts, fitness, and tiny homes. In the next section, we explain our methodology for testing this hypothesis.

2 Methodology

We begin by explaining our methodology in broad conceptual terms; we then provide a more technical explanation for readers interested in the details.

2.1 Conceptual explanation

Where does the naive YouTube watcher end up when they let the recommender system direct their viewing? The answer to this question depends primarily on two sub-questions: (i) where do they start, and (ii) what paths are most likely to be followed from various starting points? YouTube is a *dynamical system*, which makes it a moving target for research.[16] In addition, there are far too many videos on YouTube for us to answer this question comprehensively. For these reasons, we instead select the six promising topics mentioned above and operationalize them using relevant search terms. The topics are depicted in the left-hand column of Figure 1, and the search terms associated with them are depicted in the right-hand column.

That gets us our seed terms, which function as starting points. Next, we need to examine the paths that flow from these starting points. Many prominent criticisms of the YouTube recommender system suggest that it takes users down a rabbit-hole. And the model of technological radicalization articulated by Alfano et al. (2018) presumes an iterative, path-dependent process, in which users don't necessarily immediately receive recommendations for radicalizing content but eventually get there through progressive, incremental stages.

Anecdotal reports of extremism in the recommender system likewise suggest that it proceeds in stages; for example, someone might initially view videos about jogging, then receive recommendations for videos about running, then about marathons, then about ultra-marathons. For this reason, we do not want to know simply what the most-recommended videos are given a particular search: we want to know what videos are recommended by the videos that are

[16] Dynamical systems, generally, are systems with significant feedback loops. For discussion see Alfano & Skorburg (2017), Palermos (2016), Abraham, Abraham, & Shaw (1990) and Beer (1995).

recommended at the first step, and what videos are recommended by the videos that are recommended at the second step, and so on. Collecting this data manually would be extremely time-intensive, so we use a web-crawler to systematically simulate a viewer following recommendations five layers deep, noting all of the videos recommended at each stage (further technical details below in Section 2.2).

Next, we want to know not just where one ends up when one starts in certain topics but also how bad (from an epistemic point of view) that destination is. To address this point, we manually code the 100 most-recommended clips from each topic. Since we are investigating six different topics, this means that 600 videos in total are coded. As we are specifically interested in conspiracy theories, we use a three-point scale, indicating the extent of conspiratorial claims, where videos receiving a '1' have no conspiratorial content, videos receiving a '2' contain a claim that powerful forces influence (or try to influence) the topic of the video, and videos receiving a '3' contain claims that forces influence (or try to influence) the topic of the video and *also* systematically distort evidence about their actions or existence.

Note that this way of coding the videos does not prejudge all conspiracy theories as false (Dentith 2014) or all conspiracy theorizers as irrational (Coady 2007; Levy 2019; Pigden 1995, 2015). No doubt there are cases in which powerful forces influence various areas of significant social concern and even try to cover their tracks in the process, as Oreskes & Conway (2010) document at length in *Merchants of Doubt*. However, based on previous research (e.g., Klein, Clutton, & Polito 2018; Bale 2007; Keeley 1999; Sunstein & Vermeule 2009; Cook & Lewandowsky 2016; Jern, Chang, & Kemp 2009), we have concluded that what distinguishes mild from severe conspiracy theorizing — what makes severe conspiracy theories especially recalcitrant in the face of counterevidence — is the notion that the conspirators are systematically distorting the evidence. Conspiracy theories tend to become unfalsifiable when this element is added to them because evidence that is consistent with them is seen as confirmatory while evidence that is inconsistent with them is seen as fabricated (and therefore further evidence of the existence of the conspiracy).[17] It is therefore epistemically dangerous to believe such conspiracy theories. When they are true, the believer may be epistemically well off, but when they are false, the believer may be cut off from any route back to the truth. And of course, from the inside, it is impossible to distinguish these two.

[17] It is worth registering — though it is beyond the scope of what we can do here to explore in detail — some points of connection between (i) severe conspiracies in the sense of Klein et al. (2018) and (ii) epistemic echo chambers as they have been discussed in recent work in social epistemology by Nguyen (2018; see also Jiang et al. 2019). Being in an epistemic echo chamber (for example, a religious or political cult) is to be in a social-epistemic environment in which viewpoints that run contrary to the accepted viewpoint, viz., voices from *outside* the echo chamber, are met with systematic distrust. Those who subscribe to severe conspiracy theories, likewise, are inclined to a systematic distrust of views that run contrary to the conspiracy theory, and this is the case given that part of what it is to accept a severe conspiracy is to accept that there are conspirators attempting to distort evidence against the theory. Accordingly, someone who subscribes to a severe conspiracy theory in the sense of Klein et al. (2018) will *de facto* find themselves in a specific kind of echo chamber. This is the case even though the contrary does not hold, viz., not all echo chambers involve belief in conspiracy theories.

Each video is ranked on the 1-3 scale by three independent coders. In addition to coding recommended videos on our three-point scale, we keep track of how many of the most-recommended videos associated with each topic are no longer available when the coders attempt to watch them. There are several reasons why a video might no longer be available, including innocuous ones like deletion of an account, geoblocking, or a change in privacy settings. But the most common reason for removal is when a video is flagged for violating community standards.[18] These guidelines are not uniformly enforced as they rely on users to flag content that violates community standards. We code unavailable videos not with a numeral but with ‘x’.

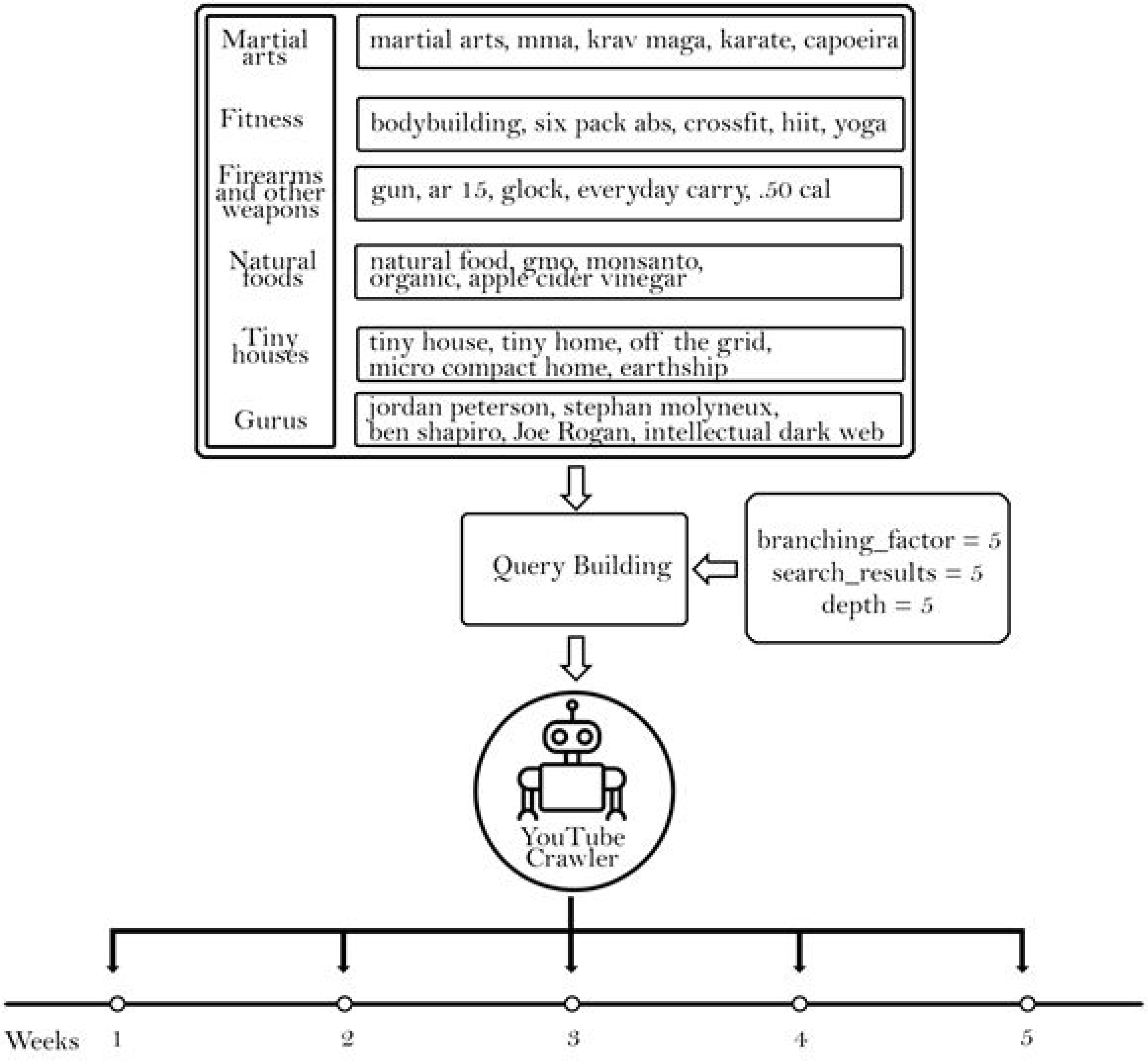
2.2 Technical explanation

On YouTube, when a user enters a search query, the YouTube search system returns the most relevant (i.e., most likely to be watched to the end by an account with this digital footprint) videos regarding the user’s query. After the user chooses one of the results, YouTube launches two separate yet closely connected operations: (i) showing the video panel and meta information, and (ii) recommending further relevant videos. When the user clicks on one of the subsequently recommended videos or lets it automatically play, the same scenario repeats, and the requested

[18] These prohibit nudity, incitement to violence or self-harm, hate speech, violent or graphic content, harassment, spam, misleading metadata, scams, threats, copyright violations, doxing, impersonation, and harm to minors. For more, see: https://www.youtube.com/about/policies/#community-guidelines.

title is displayed alongside still further recommended videos. If the video is watched to the end, the top-recommended clip is played next. Alternatively, the user may click on any of a list of recommended clips. On a typical laptop or desktop screen, five or six recommendations are visible without scrolling.

Figure 1: The schematic flow of data collection.

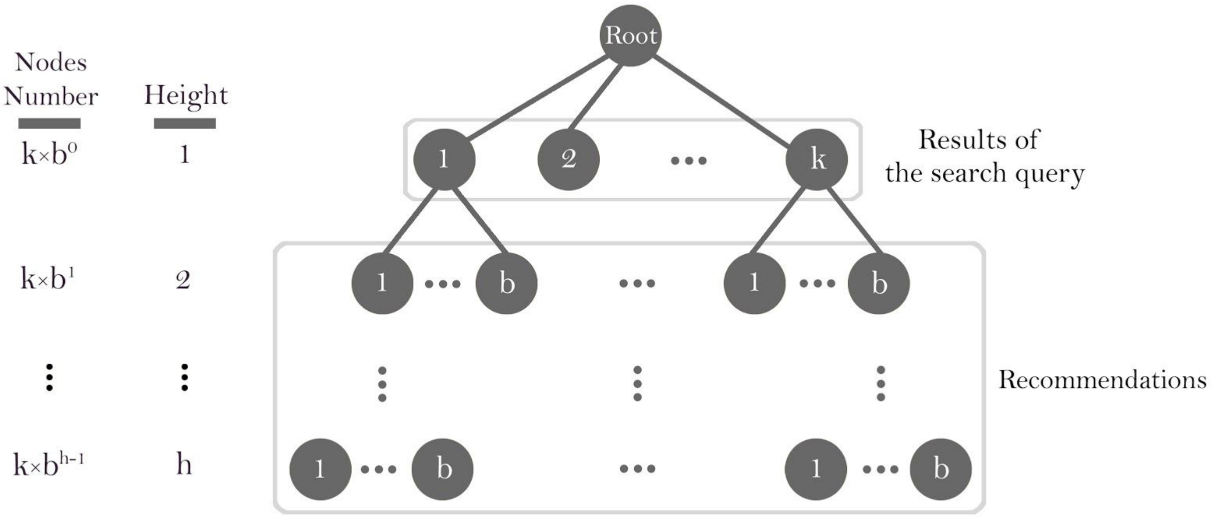

Martial arts
martial arts, mma, krav maga, karate, capoeira
Fitness
bodybuilding, six pack abs, crossfit, hiit, yoga
Firearms and other weapons
gun, ar 15, glock, everyday carry, .50 cal
Natural foods
natural food, gmo, monsanto, organic, apple cider vinegar
Tiny houses
tiny house, tiny home, off the grid, micro compact home, earthship
Gurus
jordan peterson, stephan molyneux, ben shapiro, Joe Rogan, intellectual dark web
Query Building
branching_factor = 5
search_results = 5
depth = 5
YouTube Crawler
Weeks
1
2
3
4
5

In 2016 an ex-YouTube engineer Guillame Chaslot developed a program to investigate the YouTube recommendation system during the U.S. presidential election 2016 (Chaslot 2016). This program simulates the behaviour of users on the YouTube platform by starting from a given search query and going through the recommended videos recursively until a predetermined level has been reached. The program has two modules: the crawler and the analyser. The crawler module collects the search results and recommended videos from YouTube, and the analyser stores, ranks, and visualises the results.

Figure 1 gives a schematic version of the crawler's operation, which is analogous to a *breadth-first search* (BFS). First, the user initializes the crawler by providing the search query (q), the number of search results from the search query to begin with (k), the branching factor (b), and the depth of the exploration (h). Figure 2 illustrates a simplistic version of the crawler algorithm where there are no duplicate recommendations. In such a situation, the robot starts with search query (q) results and obtains the first k videos that the YouTube search engine returns in response to the search query. Then, for every one of those videos, the robot collects the recommended videos and selects the top b recommendations recursively until it reaches the desired depth. In this case, the robot collects $S = \sum_{i=1}^{h} kb^{i-1}$ videos, including k initial videos from the search query and $S - k$ distinct recommended videos. In reality, many of the videos suggested by the YouTube recommendation system are the same, which makes the recommendation structure a potentially *cyclic* directed graph rather than a tree. In such a structure the number of recommended videos is $S < \sum_{i=1}^{h} kb^{i-1} - k = k\sum_{i=2}^{h} b^{i-1}$. After reaching the h^th^ level, the crawler module stops and the analyser module takes control. This module receives

URIs for all the collected videos as well as their corresponding metadata and stores them in a predefined path.

Figure 2: The YouTube recommendation tree when all the recommendations are distinct. In the case of the same recommended videos, the structure will be a directed graph.



For this project, we operationalize each of six topics with five search terms, meaning that we have a total of thirty seed searches. Then we launch the crawler five times for each seed search. The data collection took place over five weeks between August and September 2019.

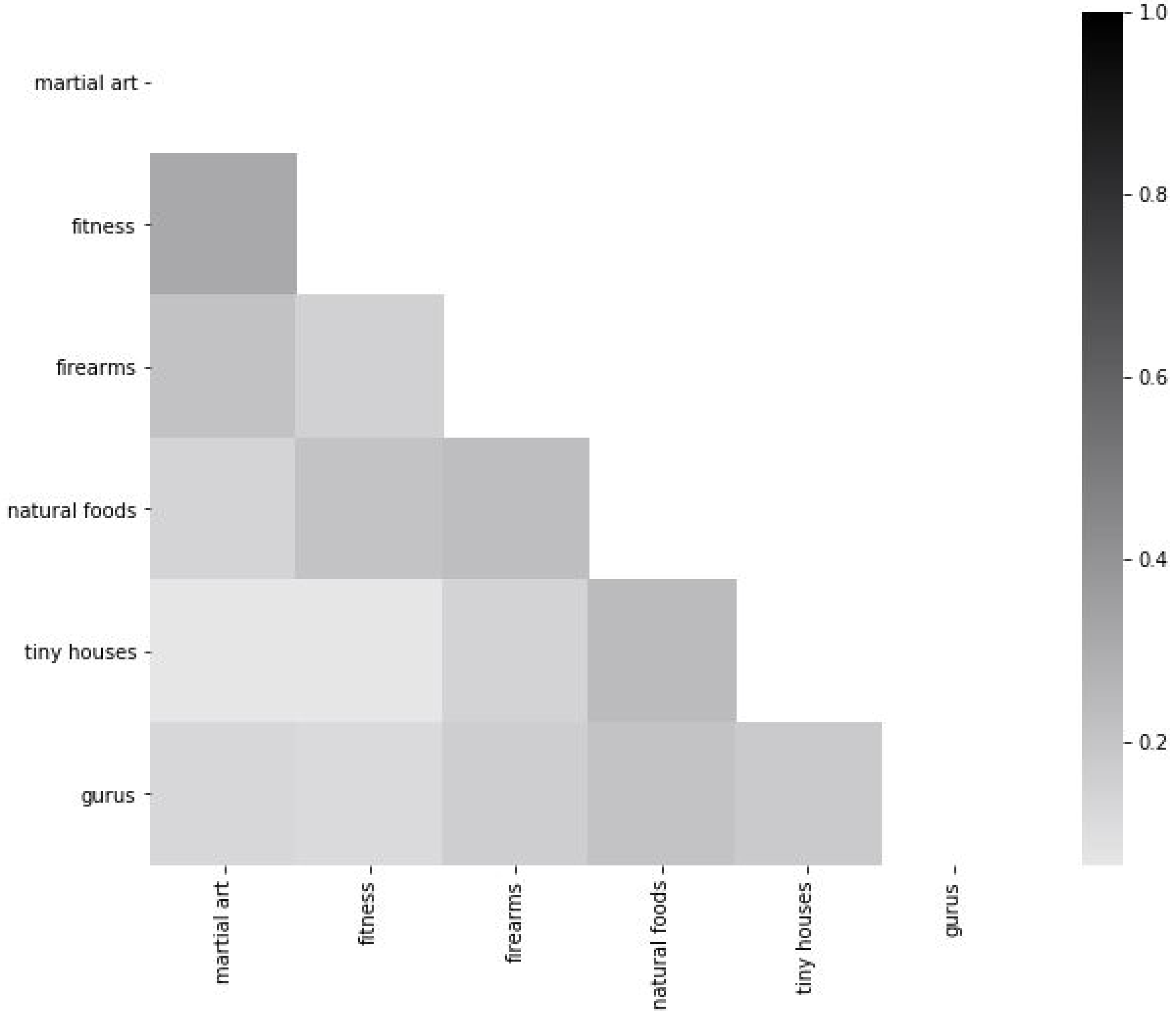
To more faithfully replicate the conditions of someone like Buckey Wolfe, we also used a virtual private network (VPN) to simulate the searches and recommendations as if the user were based in the United States (in particular, in St. Louis, Missouri). The initial arguments are set to k=5, b=5, and h=5. Each progressive stage of the process thus increases the number of returned videos by a factor of five, meaning that — for each search term — we end up collecting information about $5 + 5^2 + 5^3 + 5^4 + 5^5 = 3905$ videos. Since there are five search terms associated with each topic, this means we collect information about 19,525 URIs associated with each topic, for a total of 117,150.

Naturally, this is too many videos to code by hand, especially given that many of them are as long as three hours or more. For this reason, three independent coders evaluate the 100 most-recommended videos for each topic (600 videos total), assessing them on the three-point scale described above. In this context, most-recommended status is determined by calculating PageRankIn for each clip (Brin & Page 1998). PageRankIn represents the probability of landing on a particular node by following a random walk through the network, which means that it identifies the basins of attraction in the network of recommendations.

3 Results

Figure 3 is a similarity matrix that represents the overlap of recommendations across topics. This matrix is created based on the Jaccard similarity index. To calculate this metric, for every pair of the topics, the number of common videos in both topics is divided by all the videos associated with those topics (i.e., the intersection is divided by the union). The darker the box, the more overlap. Martial arts is more associated with fitness and firearms than with natural foods, tiny houses, or gurus. Fitness is most associated with natural foods. Firearms are somewhat oddly associated most with natural foods. Natural foods are most associated with tiny houses, and vice-versa. Gurus are somewhat more associated with natural foods than the other topics. This may be due to the fact that Jordan Peterson tends to promote his medically contentious diet of eating only red meat.[19]

Figure 3: similarity of recommendations across topics



[19] See https://www.theatlantic.com/health/archive/2018/08/the-peterson-family-meat-cleanse/567613/.

Figure 4 represents the similarity of recommendations across search terms, which suggests that some of our topics are more internally consistent than others.[20] Tiny homes and gurus are the most internally consistent, followed by firearms.

Figure 4: similarity of recommendations across search terms

[20] This matrix is created using the same method the same as the previous one. The only difference is the unit of analysis, which is more fine-grained in Figure 4. Here, instead of calculating the similarity between every pair of topics, we perform all the calculations at search-term level.

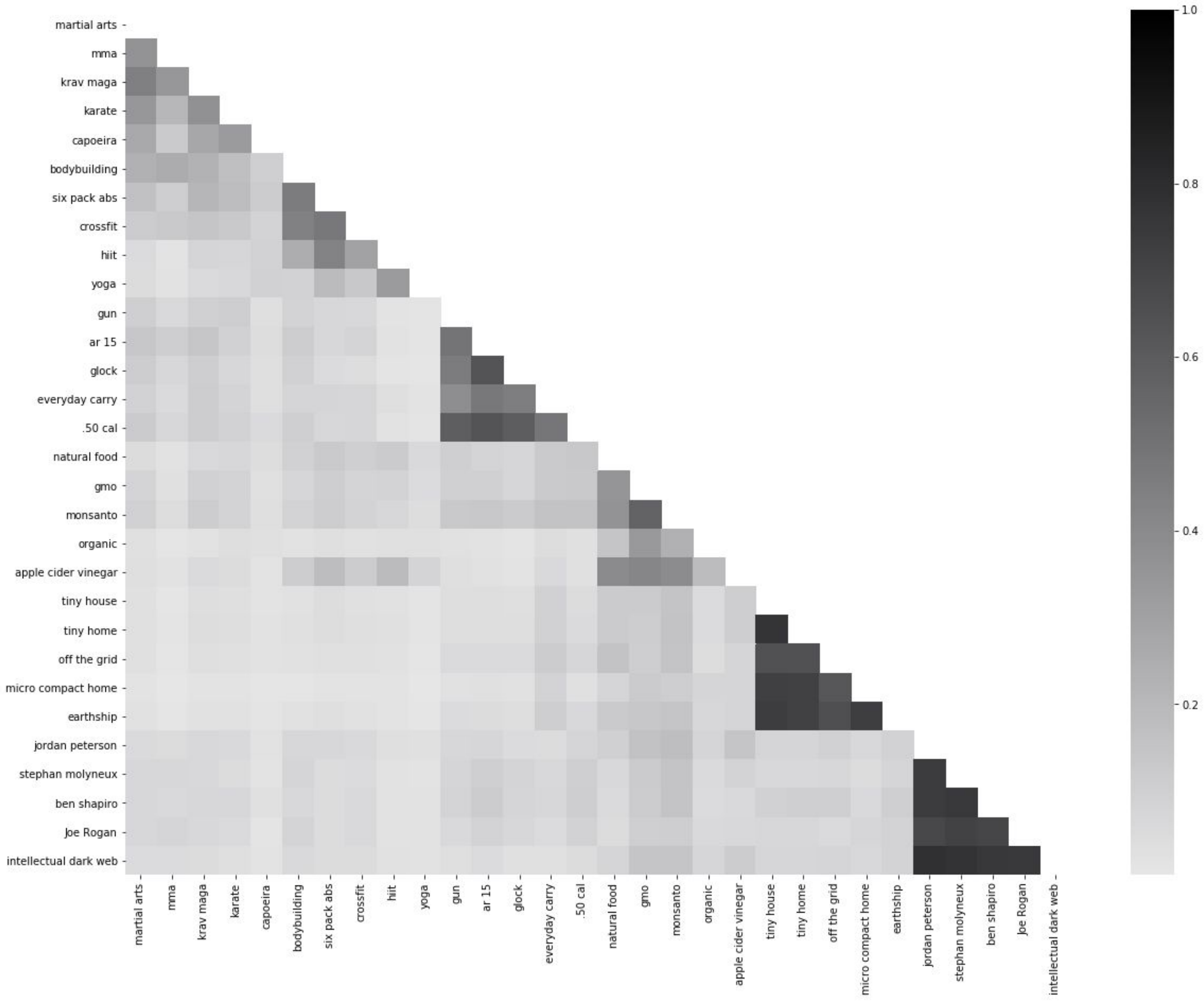


Table 1 lists some representative titles among the most-recommended clips in each category. As these examples indicate, many of the most highly-recommended clips have sensationalizing titles that make use of all-caps, exclamation points, and other standard clickbait devices.[21]

[21] For instance, they tease a revelation without giving enough details to form reasonable expectations. Which three common mistakes are made in street fights? What is the secret to mastering a handgun? What strange secret, Earl Nightingale? YouTube content creators share YouTube's interest in selling advertisements, so it is unsurprising that some of them are desperate to draw attention and curiosity with their video titles.

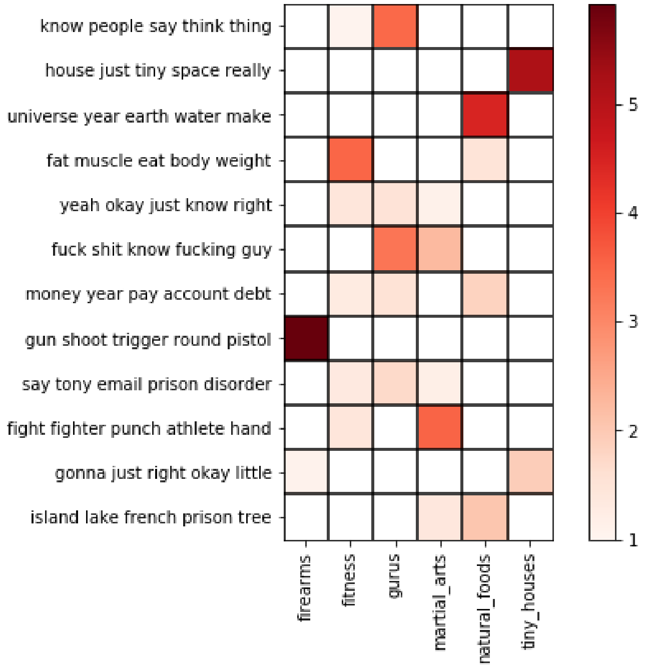
Table 1: representative titles from all six categories

topic	example titles
martial arts	20 MOST EMBARRASSING MOMENTS IN SPORTS 3 Common Mistakes in a Street Fight - Bruce Lee’s Jeet Kune Do The Gracie UFC Conspiracy
fitness	WE TRIED KETO for 45 Days, Here’s What Happened The ONLY 3 Chest Exercises You Need for MASS (According to Science) The mathematics of weight loss | Ruben Meerman | TEDxQUT (edited version)
firearms	Improvised Suppressors for .22 Rimfire 223 -vs- 5.56: FACTS and MYTHS The Secret to Mastering a Handgun
natural foods	Strange answers to the psychopath test | Jon Ronson Interstellar Travel: Approaching Light Speed The Revelation of the Pyramids (Documentary)
tiny houses	The basics on a Speed square Off-Grid Tiny House TOUR: Fy Nyth Nestled in Wyoming Mountains Surprise! Awesome figured maple (I DID NOT EXPECT THIS!!!)
gurus	Jordan B. Peterson | Full interview | SVT/TV 2/Skavlan Proven Biblical Money Principles - Dave Ramsey The Strangest Secret in the World by Earl Nightingale full 1950

We now turn from the full dataset to the 100 most-recommended clips for each topic. As an exploratory step, we built a topic model on the transcripts for the coded videos. Transcripts were retrieved using the YouTube API. Of 600 videos, 480 had transcripts available (some auto-generated, some user-entered). Transcripts were preprocessed to remove non-alphabetic material, common English stop words, words fewer than 3 characters, and descriptions of on-screen text. The resulting transcripts were then lemmatized using `nltk` (Bird, Loper, and Klein 2009). Lemmatized transcripts were transformed into a tf/idf representation (min_df=0.05, max_df=0.96), and a range of topic models were built using non-negative matrix factorization (NMF), one transcript per document. As this was an exploratory analysis on a relatively small number of documents, a 12-topic model was chosen by manual inspection as the solution that maximized discriminability while minimizing intruders.

Table 2 presents the results of the topic model. On the right are longer representations of each topic. The heatmap on the left shows, for each pair of topic and group, the percentage of transcripts that had that topic as their maximum normalized loading compared to the overall percentage of documents that had that topic as the maximized loading (ratios below 1 are cut off to improve visibility). Intuitively, this shows the extent to which a topic is over-represented in a group relative to the whole set of transcripts.

Table 2. Left: p(Max | Group) / p(Max), bounded at 1. Y-axis shows top 5 topic words. Right: Longer representations of topics.



0 know people say think thing mean talk want life right question good way lie world person really woman make just conversation child idea tell time
1 house just tiny space really storage water build home kind window thing wall want little use door bed actually lot design kitchen wood love work
2 universe year earth water make planet star time light energy ancient solar theory scientist use world hole space sun discover large wave surface
3 fat muscle eat body weight exercise calorie food fast day protein diet energy workout meal sugar burn want time store week carbon really make
4 yeah okay just know right think really gonna good yes wow challenge coin mean bike come glass pretty look create guy sing feel win hmm
5 fuck shit know fucking guy just people man gonna yeah right say mean dude thing think like work want time come everybody talk movie play
6 money year pay account debt number dollar make bank million tax say know cash life game business rich check kid day card family work note
7 gun shoot trigger round pistol rifle bullet barrel carry shot target grip weapon magazine millimeter just load hand come inch use pull knife
8 say tony email prison disorder send tell know list bank look think disease mental year man yes people address claim gate come brian story phone
9 fight fighter punch athlete hand win come round right opponent shot time just score knock pound hook sport ring ball game hit left goal player
10 gonna just right okay little let want look bit guy good know card cut thing gold really start kind come piece way trick number try
11 island lake french prison tree mystery ocean float mysterious site map sea number escape mexico century home specie exist day france base build

Some of the results are unsurprising. Firearms, fitness, martial arts, and tiny houses each have a unique characteristic topic, one which loads on words that one would expect from those videos. This shows that the topic model was able to extract sensible patterns from the data. Natural foods also has a unique high-loading topic, but one which appears to emphasize a common core of fringe scientific ideas. This suggests that the popular videos in natural foods are not unified by their particular recommendations so much as their adherence to a loose set of beliefs that are used to justify their content.

The pattern seen with the guru videos differs from that of the other five. The two topics that guru videos load most heavily on are “rhetorical” topics which are characterized by a manner of speaking — one more congenial, the other more angry. In other words, what appears to be most characteristic of guru videos is not a specific content but a more general manner of speaking. Insofar as there are similarities, they are mostly with the fitness and martial arts categories, suggesting perhaps a rhetorical style more broadly associated with an exaggerated masculinity. This exploratory may be worth further investigation.

Table 3 provides summary details for each of the 100 most-recommended clips. The most-viewed topic was martial arts, followed by natural foods, fitness, gurus, firearms, and tiny houses. The most-liked topic was natural foods, followed by martial arts, gurus, fitness, firearms, and tiny houses. The most disliked topic was martial arts, followed by natural foods, gurus, fitness, firearms, and tiny houses. The longest videos were associated with gurus, followed by tiny houses, fitness, natural foods, firearms, and martial arts.

Table 3: summary statistics for each topic. All figures report averages (means)

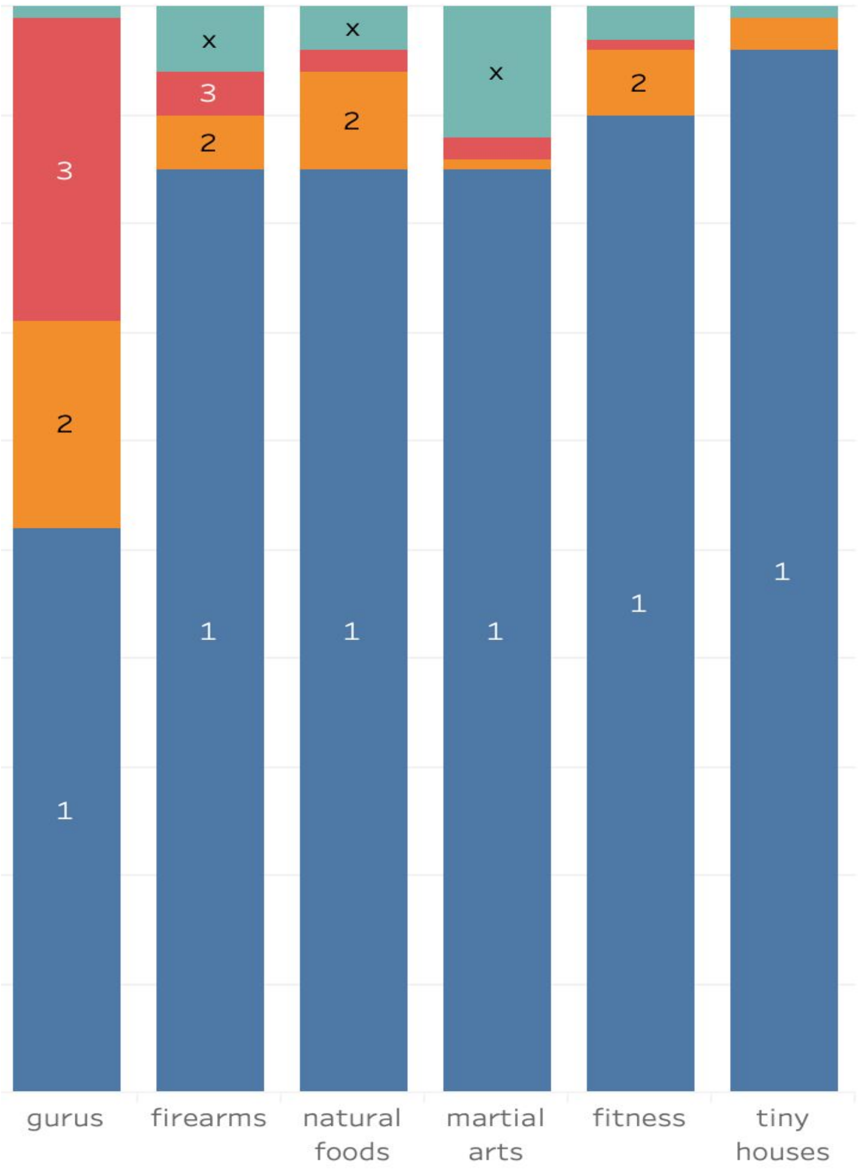
category	views	likes	dislikes	length in seconds
martial arts	11,094,353	81,015	6,381	908.3
fitness	6,029,494	67,272	3,883	1409.94
firearms	4,746,370	41,300	3,086	1185.96
natural foods	6,727,368	96,167	4,939	1253.47
tiny houses	3,505,384	40,158	2,348	1619.3

gurus	4,839,331	70,172	3,939	3398.25

Each of these 600 clips was independently coded by three different coders according to the scheme described above. We observed adequate interrater reliability (Fleiss's $\kappa = .445$, $z = 27.5$, $p < .0001$). To arrive at finalized ratings, we used the following decision procedure. First, if all three raters agreed, then their consensus was entered as the final rating of the clip. Second, if two of three raters agreed but the third disagreed, then we entered the value agreed-upon by the majority as the final rating of the clip. Finally, if all three raters disagreed (meaning that the clip received scores of 1, 2, and 3), one member of the research team reviewed the clip a second time and came to a final conclusion. Such maximal disagreement occurred in just 14 out of 600 cases (2.3%).

Figure 5 represents the severity of conspiracy theories among the 100 most-recommended clips from each topic.

Figure 5: distribution of conspiracy theories among the most-recommended clips from each topic. 1 = no conspiracy theory, 2 = mild conspiracy theory, 3 = severe conspiracy theory, x = clip no longer available at time of coding



As Figure 5 makes clear, the YouTube recommender system does indeed promote conspiracy theories from all six topics.[22] However, the proportion and severity differ from topic to topic. To

[22] Does the recommender system promote *more* conspiracy theories than some sort of neutral baseline? We are unable to address this question in the current study because we have no way of ascertaining what a neutral baseline might be. It might be possible to compare one recommender

test our main hypothesis, we calculate for each topic the ratio of the number of clips that received a rating of 2 or 3 to the number of clips that received a rating of 1, 2, or 3. This leaves out clips that received a rating of x. This analytic method may miss some information, but because the list of reasons that a clip might be unavailable is so diverse, we decided not to presume that unavailable clips were or were not conspiratorial. The resulting ratios are represented in Table 4.

Table 4: ratio of conspiratorial clips to rated clips for each category

topic	**ratio**
gurus	.475
natural foods	.115
firearms	.096
fitness	.072
martial arts	.034
tiny houses	.030

Remarkably, *nearly half* of the visible most-recommended videos from the gurus topic were conspiratorial. The other topics seem less worrisome, though still problematic. Over 10% of the visible most-recommended videos from the natural foods topic were conspiratorial, as were

system to another, or to compare this recommender system to an older version of the same recommender system. However, we lack access to these comparators. What we *have* established is that the YouTube recommender system does in fact push conspiracy theories, not that it pushes them harder than they would be pushed by an alternative. Thanks to an anonymous reviewer for raising this point.

nearly 10% of the videos from the firearms topic. Thus, the most conspiracy-heavy topic by far was associated with the political right, and the next two were split between natural foods and the right firearms. It is fairly clear that firearms are associated with the political right; natural foods might seem like a left-wing interest but is politically ambiguous, as the film *Dr. Strangelove* illustrates with its gag about "precious bodily fluids."

Recall that our pre-registered hypothesis was that the proportion of conspiracy theories associated with different topics would be ordered as follows: gurus = firearms > natural foods > martial arts > fitness > tiny houses. This hypothesis is largely borne out. The actual ordering is gurus > natural foods > firearms > fitness > martial arts > tiny houses. In other words, the second- and third-ranked items as predicted turned out to be the third- and second-ranked items in the actual data, while the fourth- and fifth-ranked items as predicted turned out to be the fifth- and fourth-ranked items in the actual data. The top item (gurus) and the last item (tiny houses) were correctly predicted.

As an exploratory analysis, we also examined the clips with each rating to see which stage of data collection they cropped up in. Figure 6 shows these results.

Figure 6: fraction of top-recommended videos discovered at each stage of data collection.

Figure 6a: videos labeled 1

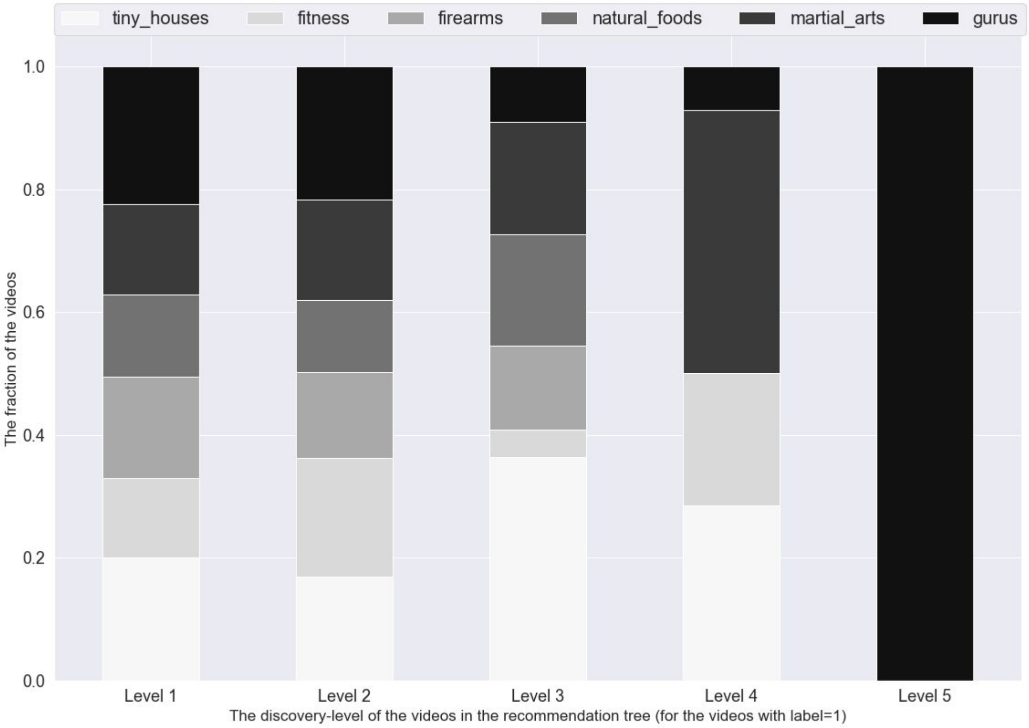


Figure 6b: videos labeled 2

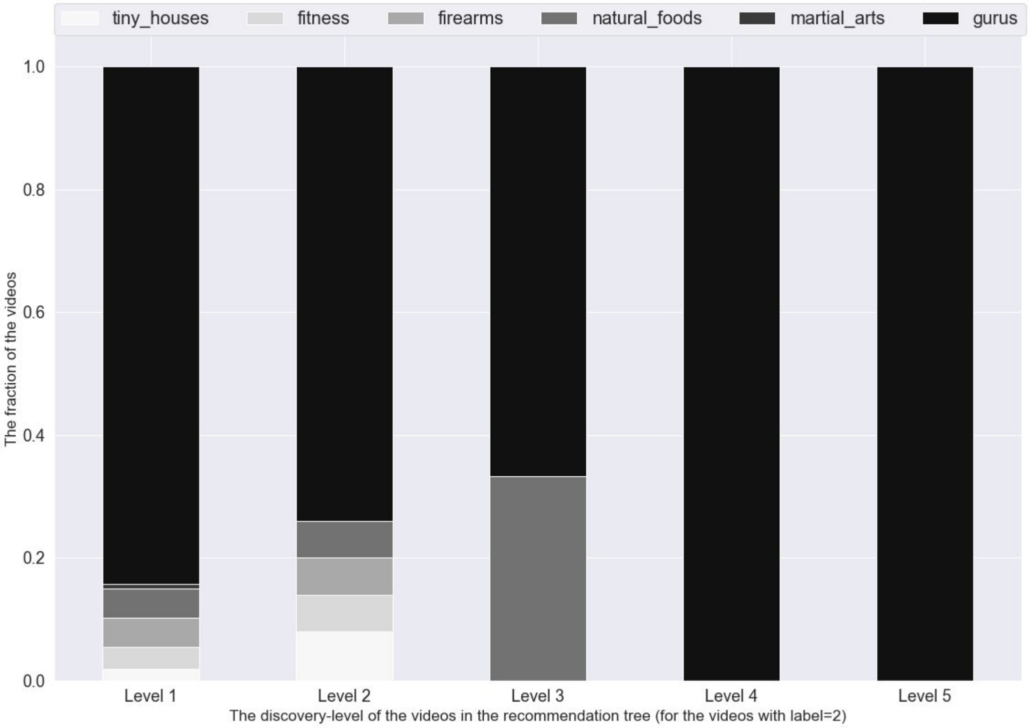


Figure 6c: videos labeled 3

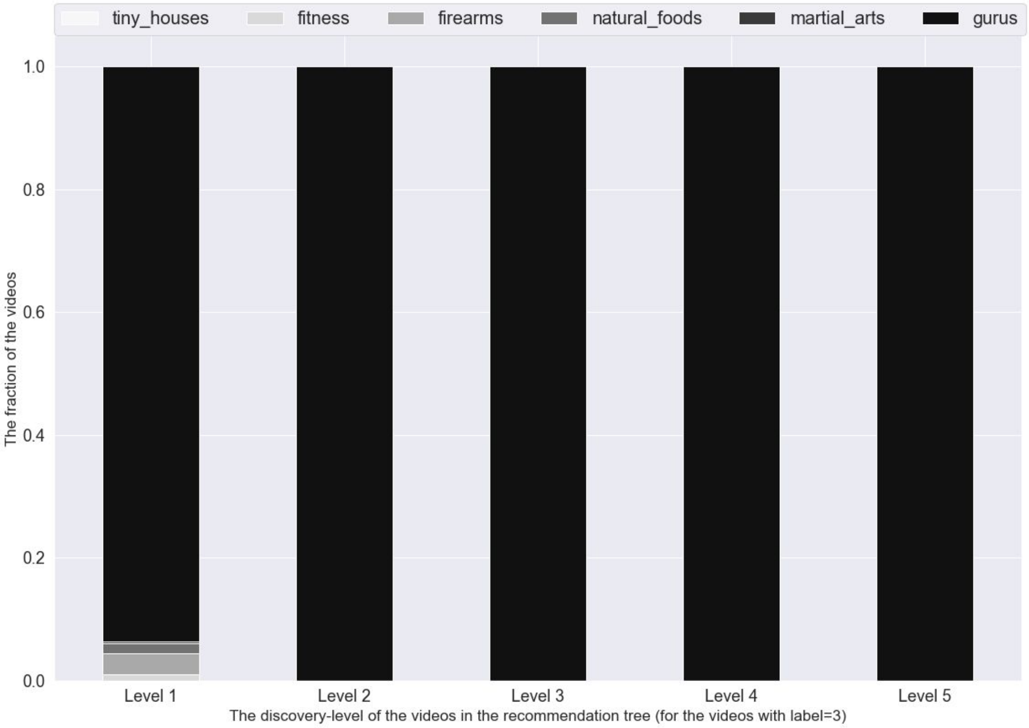
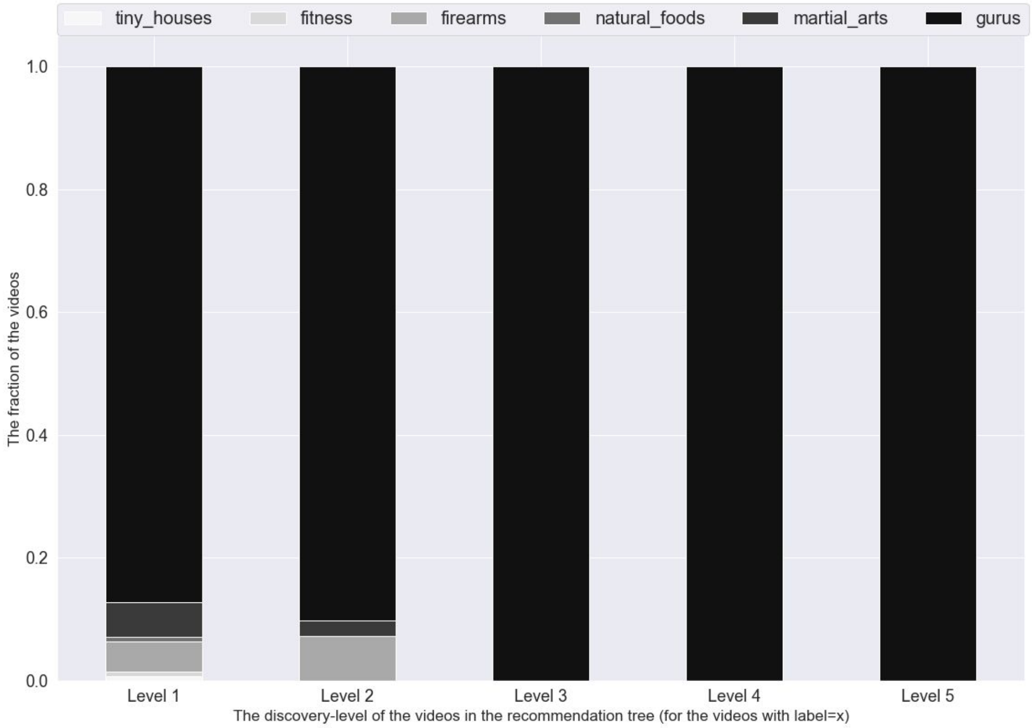


Figure 6d: videos labeled x



As figure 6 shows, conspiratorial content can appear in any depth of the recommendation tree/graph. For some topics, they only appear in early levels, while for other topics conspiratorial contents keep showing up in deeper levels. Second, between mild and severe conspiracy theories, the earlier has a higher chance of appearance in deeper levels of the recommendation tree/graph. In 6c, there are severe conspiracy theories from different topics only in the first level (and in other levels the conspiracy theories belong to one particular topic). By contrast, in 6b, mild conspiratorial content belonging to different topics are available in the first three levels. Third, as we can observe in Figure 6a, until level 4 videos are coming from 4 different categories, which means even in deep levels of recommendation tree/graph we should expect to see non-conspiratorial contents. Finally, in general, the guru videos tend to be recommended more the deeper one goes into the tree. Level 5 (the final bar of each histogram) is almos all gurus for every single label. This suggests that the "rabbit hole" effect is especially pronounced in the case of guru videos.

4 Discussion, future directions, and recommendations

In this study, we used data-mining and expert coding to test the technological seduction hypothesis, according to which the YouTube recommender system is liable to promote and amplify conspiracy theories, and that it does so differentially depending on the topic that first brings users to the platform.

4.1 Discussion

While we motivated our study with the extreme case of Buckey Wolfe, we should emphasize that the promotion of conspiracy theories need not be extreme to be harmful. Numerous studies have associated conspiracy theorizing with the rejection of authoritative scientific findings (Simmons & Parsons 2005; Bogart and Thorburn 2005; Jolley & Douglas 2014a; Lewandowsky et al. 2013), with willingness to eschew medically sound procedures like vaccination (Jolley & Douglas 2014b; Oliver & Wood 2014; Dunn et al. 2015), and with general disengagement from mainstream political and social realms (Jolley & 2014b). Furthermore, studies of online conspiracy theorizing emphasize the broad scope that online platforms can have for influence and participation among otherwise uncommitted individuals (Dunn et al. 2015; Klein et al.

2018). Given the extensive reach of recommender systems — with YouTube alone reaching two billion users per month who watch over one billion hours of video in total per day[23] — one would expect even minor effects to have substantial consequences in a large population over a longitudinal timescale.

As we noted above, YouTube's recommender system is itself a moving target — indeed, it has attracted attention precisely because a change to the algorithm appears to have shifted the balance towards promoting longer, more conspiratorial content. Our research supports that claim. Yet YouTube is constantly tweaking its algorithm (making replication and reproducibility of work like ours tricky), and content-providers constantly tweak their output in order to maximize views within the system. Our research thus represents a snapshot of a complex, evolving system. Even if the YouTube recommender algorithm changes to avoid these negative consequences (and we rather hope that it does![24]), the present work represents a valuable picture of how

[23] See https://www.youtube.com/about/press/.

[24] A qualified note of optimism here owes to recent efforts by YouTube's parent company Alphabet to limit the demographics of election ad targeting to the three general demographics of age, gender, and location. This new implementation is scheduled to go into effect in January 2020, and it reflects an (albeit minimal) effort on the part of Google to disrupt the way potentially misleading information is disseminated. See https://www.blog.google/technology/ads/update-our-political-ads-policy/. However, there is perhaps more cause for pessimism about what to expect from further tweaks to the recommender system. Insofar as the algorithm's design function continues to aim at maximising watch-time within the system, 'improvements' to this design in the form of further tweaks are (provided they

epistemologically problematic beliefs and belief-forming processes can be encouraged and strengthened by apparently innocuous algorithmic decisions.

Our seed terms were designed to capture an array of potential conspiracy theories. The strikingly higher proportion of guru videos with conspiratorial content is notable. These videos tended to have a substantial political aspect, and the conspiracy theories they endorse interact with that content. This is interesting in light of accounts that emphasize the political function of conspiracy theorizing. Hofstadter (1964)'s classic analysis noted the attraction of conspiracy theories to those who saw only the effects of political power, not its inner workings.[25]

Other accounts have emphasized the link between conspiracy theories and perceived powerlessness (Abalakina-Paap et al. 1999; Whitson & Galinsky 2008), lack of trust in authority (Swami et al. 2010), and the active desire to display defiance towards established norms and institutions (Goezel 1994). While these accounts rarely defend conspiracy theorizing as such, some do note that conspiracy theories play a role in making salient the complaints of marginalized groups (Miller 2002) or drawing attention to historical inequalities (Thomas and Quinn 1991; Bogart and Thorburn 2005). However, conspiracy theories may play a politically

are effective) going to be tweaks that only *further* serve to recommend the very kinds of conspiratorial content that is likely to hold attention. With this in mind, the fact that YouTube's recommender system is a 'moving target' not only makes it difficult to study, but also gives it the potential to transform (absent further regulation) into an even more epistemically pernicious mechanism of technological seduction than it is presently.

[25] For additional recent discussion about the relationship between conspiracy theories and political ideology, see Cassam (2019).

convenient role for those in power, by distracting from and thereby masking underlying inequalities. Cassam (2019) suggests that conspiracy theories are forms of propaganda with a knowledge-destroying political function.

4.2 Future directions

We should note a significant limitation of this study: namely, that we only used one exploratory account, and it had no history, no prior digital footprint. This is the equivalent of someone first using YouTube before they use *any* Alphabet service or other service that provides data to Alphabet. The recommender system could only personalize based on the sixty initial seed searches and the geographic location of the VPN. In other words, it could not personalize based on prior YouTube viewing, general search (Google Search), location history (Google Maps), correspondence (Gmail), social networking (G+), translation requests (Google Translate), access to books (Google Books), authorship of or interaction with scholarship (Google Scholar), ownership and usage of Android devices, and whatever else Alphabet has access to, which probably includes medical history, police record, voter registration record, and more. That is naturally not how the vast majority of people first encounter YouTube. In defense of this limitation, we should note that our study was designed to emulate an ideally naïve, passive agent who was entirely at the mercy of the recommender system. Ordinary people disengage with content they find intrinsically unpalatable. Conversely, individuals who do end up engaging with conspiracy theories online likely do so through some mix of technological seduction and intrinsic interest in conspiracy-adjacent topics (Klein, Clutton, & Dunn 2019). Nevertheless, this idealization reveals how technological seduction can have a profound effect even *without* active intervention on the part of the user.

Future research could address this shortcoming by creating a small army of dummy accounts, building a diverse range of digital footprints for them, and then letting them interact with the YouTube recommender system. Additional future research could address not just the YouTube platform but other popular platforms that also have the potential to promote and amplify conspiracy theories. It would be especially interesting to see whether some platforms do a better job than others of tamping down conspiracy theories, and to examine what features of those platforms account for their success.

Further directions for research include moving from how-possibly to how-actually explanation. In this paper, we merely show that there is a robust pathway from some seemingly anodyne topics to outright conspiracy theories. The how-possibly explanation detailed here, to be sure, is worrying in what it illuminates; it can help us to make straightforward sense of how (as the technological seduction thesis predicts) individuals can become self-radicalized relatively easily, e.g., with a minimal kind of online friction. And further, it helps us to make sense of how such radicalization can occur without an initial intent to even consider extreme views, and in the absence of traditional kinds of social contacts with like-minded thinkers.

However, to be clear, we do not establish that actual users follow the particular pathway we detail, nor do we furnish evidence of how many of them might do so. Perhaps most importantly, we have yet to provide evidence that the recommender system leads to transformative experiences of the sort that might be necessary to bring about a significant kind of ‘shift’ (for better or worse) in one’s epistemic perspective.[26] It might instead be that users who are already inclined to accept conspiracy theories fill in the details with content from YouTube, not that YouTube takes non-conspiracy theorizers and turns them into conspiracy theorizers. To establish any of these stronger hypotheses, it would be necessary to randomize users (or at least pseudo-users) to engage with various topics and measure what happened to their acceptance of relevant conspiracy theories. While such research would be fascinating, conducting it without violating constraints on research ethics would be challenging.

Finally, it would be valuable to investigate predictors of conspiracy-theorizing in YouTube clips. Potential predictors include the channel that posts the clip, the title of the clip, the transcript of the clip, the length of the clip, and which other clips point to the clip via the recommender system. Investigating this at scale might make it possible to flag potentially problematic videos automatically and shunt them to human classifiers for further review.

4.3 Recommendations

[26] The idea that such shifts might be epistemically important is key to Paul’s (2014) influential work on the epistemology of transformative experience. According to Paul, the adoption of certain kinds of perspectives requires a significant experience, one that won’t necessarily be secured by an incremental exposure to a certain kind of evidence (or apparent evidence) in favour of that perspective. This gloss of Paul’s view is of course compatible with there being some possible cases where incremental change can elicit a transformative experience; her position does not foreclose that possibility.

We conclude by offering some recommendations that seem apropos in light of the current study. We begin with recommendations to users of systems like YouTube, followed by recommendations to owners and developers of such platforms, and finally recommendations to policymakers and regulators.

One natural recommendation for users is to develop the dispositions or epistemic virtues that answer to the new online epistemic environment in which they find themselves. If Foot (1997, p. 3) is right in thinking that "virtues are in general beneficial characteristics, and indeed ones that a human being needs to have, for his own sake and that of his fellows," then when our epistemic needs change, so do the dispositions that answer to them. Some recent work in virtue epistemology emphasizes the need to rethink the virtues in light of our evolving, digitized epistemic environment (Vallor 2016). For example, Alfano & Klein (2019) point out that the Internet has catalyzed both quantitative and qualitative shifts in the information ecology along multiple dimensions:

Volume: we have access to more information.

Velocity: we have access to information more quickly and fluently.

Veracity: we have access to more accurate information.

Variety: we have access to more diverse information sources.

Voice: we have more power to make ourselves and others heard.

In the case of conspiracy theories being promoted by YouTube, the problem seems to center on volume, velocity, and veracity: streaming content never stops coming in, especially when we allow YouTube to autoplay suggestions. Moreover, while there are many accurate sources to be found even on YouTube, there are also many inaccurate and conspiratorial sources, and they seem to be amplified by the recommender system.

Users ought to exercise *epistemic vigilance* in order to avoid being seduced by such conspiracy theories (Sperber et al. 2010)). The exercise of vigilance can of course be an ongoing, conscious, intentional activity (what is typically considered a virtuous disposition), but it could also involve automating or *de*-automating certain behaviors. For instance, users could simply switch off the autoplay default, which would force them to actively choose whether to watch another video, as well as to choose which video that would be. Other authors emphasize a range of additional epistemic virtues that may come in handy when dealing with our new online epistemic environment. For example, Heersmink (2018) addresses curiosity, intellectual autonomy, intellectual humility, attentiveness, intellectual carefulness, intellectual thoroughness, open-mindedness, intellectual courage, and intellectual tenacity. Many of these dispositions seem relevant on their face to avoiding being sucked into conspiracy theories by the YouTube recommender system and other online tools. Indeed, Meyer (2019) has shown that intellectual humility as measured by the scale introduced in Alfano et al. (2017) predicts acceptance and rejection of both fake news and conspiracy theories furnished by online sources: participants who score higher in intellectual humility tend not to be taken in.

In addition to users who only consume content on YouTube, there are content creators. They face an incentive structure that induces them to produce problematic content insofar as they want to maximize the number of users who view and like their videos. Moreover, content creators whose accounts are monetized share the same incentives as YouTube itself: the more time people spend watching their clips, the more money they make from advertisements. In other words, YouTube actively incentivizes the production of exactly the sort of problematic videos that this study has investigated. Nevertheless, we believe that content creators share some (though clearly not all) responsibility for what they do, so we suggest that they ought not game the recommender system by publishing conspiracy theories. Likewise, experts such as academics may have an imperfect duty to (help to) create content that competes with conspiracy theories. Among the most-recommended clips that we investigated, there were quite a few TED talks by academics and other experts. While we have our reservations about the depth of TED talks, we much prefer them to the conspiracy theories of Jordan Peterson and Ben Shapiro.

Of course, individuals on their own are not necessarily well-situated to address the influence of vast, powerful, and wealthy corporations like Alphabet (and its subsidiaries like YouTube). Workers, managers, and owners in the technology industry should hold themselves to higher standards than they currently do. For instance, YouTube should be willing to sacrifice some profit margin to address the sort of problem discussed in this paper. Optimizing for watch-through without attending to and mitigating potential deleterious side-effects such as the promotion of conspiracy theories and corresponding radicalization of users is myopic and almost psychopathically greedy.[27] If YouTube and other relevant companies (e.g., Facebook, Twitter)

[27] The monetization of conspiracy theories is, of course, not limited to YouTube, which has been our focus, and the fact that conspiracy theories are monetizable is well established. Sunstein (2014) refers to those who engage in the wider practice of profiting off of conspiracy theories 'conspiracy entrepreneurs' (2014: 12), a classic example of which he offers is Alex Jones of InfoWars. It is worth noting the important gap between *conspiracy theorising* and *conspiracy entrepreneurship*. Though Jones (like YouTube) profits from the production of conspiratorial

were to take seriously the influence and power that they wield, they would also provide resources to independent researchers to investigate the effects of that power. Such resources include not just money but also access to data and to internal developers, who could be interviewed by qualified social scientists to help understand how the company's algorithms were developed and how they work. YouTube could also more stringently enforce community standards and invest heavily in independent fact-checking, which would enable the platform to take down extremely problematic content and avoid promoting (through recommendations and monetization) less problematic content. Finally, YouTube could simply stop making autoplay the default setting. As we know from much social scientific research, defaults are extremely powerful, so a tweak as simple as this could make a big difference (Sunstein 2013). Doing so might be part of a larger effort to get away from the advertisement-based business model that made watch-through such a tempting target for optimization in the first place, which would have the additional benefit of reducing YouTube's reliance on surveillance capitalism.

content, publicly available court documents cast doubt on whether he himself believes the content he profits from. See https://www.theguardian.com/us-news/2019/mar/30/alex-jones-sandy-hook-claims-psychosis. YouTube, being a large corporation, presumably does not have intentional states such as beliefs. It is also worth highlighting the important gap — vis-a-vis YouTube — between the the epistemic badness of (i) believing conspiracy theories; and (ii) facilitating the belief in such theories in others. We've demonstrated how YouTube's recommender system can easily bring about the second kind of epistemic bad, but we presume that YouTube itself lacks beliefs.

Finally, regulators and policymakers have a serious role to play in addressing our brave new epistemic world. On the enforcement side, they could and should pass new rules and laws that require large, powerful corporations to provide the kind of support to independent researchers mentioned above. Such rule-making and legislating could be modeled to some extent on the European Union’s recent General Data Protection Regulation (GDPR), but would have to go a great deal further. On the investment side, governments should see to it that schools teach students digital literacy, starting at a young age and continuing into university education. This digital literacy initiative would focus on critical thinking and cultivation of the epistemic virtues discussed above.

References

Abalakina-Paap M., Stephan W. G., Craig T., & Gregory W. L. (1999) Beliefs in Conspiracies. *Political Psychology*, 20(3): 637–647.

Abraham, F. D., Abraham, R. H., & Shaw, C. D. (1990). *A visual introduction to dynamical systems theory for psychology*. Aerial Press. Retrieved from http://psycnet.apa.org/psycinfo/1991-97299-000

Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association*, 4(3): 298-322.

Alfano, M., Iurino, K., Stey, P., Robinson, B., Christen, M., Yu, F., & Lapsley, D. (2017). Development and validation of a multi-dimensional measure of intellectual humility. *PLoS ONE*, 12(8): e0182950.

Alfano, M., & Klein, C. (2019). Trust in a social and digital world. *Social Epistemology Review and Reply Collective*, 8(10): 1-8.

Alfano, M. & Skorburg, J. A. (2017). The embedded and extended character hypotheses. In J. Kiverstein (ed.), *Handbook of Philosophy of the Social Mind*. 465-78. Routledge.

Alfano, M. & Skorburg, J. A. (2018). Extended knowledge, the recognition heuristic, and epistemic injustice. In D. Pritchard, J. Kallestrup, O. Palermos, & J. A. Carter (eds.), *Extended Knowledge*. Oxford University Press.

Bale, J. M. (2007). Political paranoia v. Political realism: On distinguishing between bogus conspiracy theories and genuine conspiratorial politics. *Patterns of Prejudice*, *41*, 45–60.

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1), 173–215.

Bird, S, Loper,E., & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

Bogart L. M. & Thorburn S. (2005). Are HIV/AIDS conspiracy beliefs a barrier to hiv prevention among African Americans? *Journal of Acquired Immune Deficiency Syndromes,* 38(2): 213–218.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30: 107-17.

Cassam, Q. (2019). *Conspiracy Theories*. Polity.

Chaslot, G. (2016). *Exploring YouTube recommendations*. Available at https://github.com/pnbt/youtube-explore.

Coady, D. (2007). Are conspiracy theories irrational? *Episteme*, 4(2): 193-204.

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*, 160–179.

Dentith, M. (2014). *The Philosophy of Conspiracy Theories*. Palgrave.

Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D., & Coiera, E. (2015) Associations Between Exposure to and Expression of Negative Opinions About Human Papillomavirus Vaccines on Social Media: An Observational Study. *Journal of Medical Internet Research,* 17(6): e144

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting False Information in Memory: Manipulating the Strength of Misinformation Encoding and its Retraction. *Psychonomic Bulletin & Review,* 18: 570–578.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.

Foot, P. (1997). Virtues and vices. In R. Crisp & M. Slote (eds.), *Virtue Ethics*, pp. 163-77. Oxford University Press.

Forrester, J. (1990). *The Seductions of Psychoanalysis: Freud, Lacan, and Derrida*. Cambridge University Press.

Gigerenzer, G. (2008). *Rationality for Mortals: How People Cope with Uncertainty*. Oxford University Press.

Gigerenzer, G. & D. Goldstein. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* 103(4): 650.

Goertzel T. (1994) Belief in conspiracy theories. Political Psychology. 15: 731–742.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE,* 7(9): e45457.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B. and Johansson, P. (2013). How the Polls Can be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions. *PLoS One* 8(4): e60554.

Heersmink, R. (2017). A virtue epistemology of the internet: Search engines, intellectual virtues and education. *Social Epistemology*, 32(1).

Hofstadter R. (1964) The paranoid style in American politics. *Harper's Magazine* 229(1374): 77–86.

Icke, D. (1999). *The Biggest Secret: The Book That Will Change the World*. Bridge of Love Publications.

Jern, A., Chang, K.-m. K., & Kemp, C. (2009). Bayesian belief polarization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 853–861).

Jiang, R., Chiappa, S., Lattimore, T., Gyorgy, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI.*

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science* 310: 116–9.

Jolley D., & Douglas K. M. (2014a) The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLOS ONE,* 9(2): e89177.

Jolley D., & Douglas K. M. (2014b) The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint. *British Journal of Psychology,* 105(1): 35–56.

Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy*, *96*, 109–126.

King, O. (2019). Presumptuous aim attribution, conformity, and the ethics of artificial social cognition. *Ethics and Information Technology*.

Klein, C., Clutton, P., & Polito, V. (2018). Topic modeling reveals distinct interests within an online conspiracy forum. *Frontiers in Psychology,* 9: 189.

Klein, C., Clutton, P., & Dunn, A. (2019). Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *PLOS ONE*.14(11): 1-23.

Levy, N. (2017). The Bad News About Fake News. *Social Epistemology Review and Reply Collective,* 6(8): 20-36.

Levy, N. (2019). Is conspiracy theorising irrational? *Social Epistemology Review and Reply Collective,* 8(10): 65-76.

Lewandowsky S., Gignac G. .E, & Oberauer K. (2013). The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLOS ONE*, 8(10): e75637.

Lewandowsky, S., Kozyreva, A., & Ladyman, J. (2020). What rationality? Comment on Levy. *Social Epistemology Review and Reply Collective*, 8(10).

Miller, S. (2002). Conspiracy theories: public arguments as coded social critiques: a rhetorical analysis of the TWA flight 800 conspiracy theories. *Argumentation and Advocacy,* 39(1): 40–56.

Nguyen, C. T. (2018). Echo chambers and epistemic bubbles. *Episteme,* 1-21.

O'Connor, C. (2019). *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford University Press.

Oliver J. E., & Wood T. (2014). Medical conspiracy theories and health behaviors in the United States. *JAMA Internal Medicine*, 174(5): 817–818.

Oreskes, N. & Conway, E. (2010). *Merchants of Doubt*. Bloomsbury.

Palermos, S. O. (2016). The dynamics of group cognition. *Minds and Machines*, 26(4): 409–440.

Paul, L. (2014). *Transformative Experience*. Oxford University Press.

Pigden, C. (1995). Popper revisited, or what is wrong with conspiracy theories? *Philosophy of the Social Sciences*, 25(1): 3-34.

Pigden, C. (2015). Conspiracy theories and the conventional wisdom revisited. In O. Loukola (ed.), *Secrets and Conspiracies*. Rodopi.

Prentice, D. A., Gerrig, R. J. (1999). Exploring the Boundary Between Fiction and Reality. In Shelly Chailen & Yaacov Trope (eds.) *Dual Process Theories in Social Psychology*, 529-546. New York: Guilford Press.

Resnik, D. (1991). How-possibly explanations in biology. *Acta Biotheoretica*, 39: 141-49.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V., & Meira Jr, W. (2018). Auditing Radicalization Pathways on YouTube. In *Woodstock '18: ACM Symposium on Neural Gaze Detection*, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 18 pages. https://doi.org/ 10.1145/1122445.1122456

Simmons W. P. & Parsons S. (2005). Beliefs in conspiracy theories among African Americans: A comparison of elites and masses. *Social Science Quarterly*, 86(3): 582–598.

Simion, M., Kelp, C., and Ghijsen, H. (2016). Norms of belief. *Philosophical issues*, 26(1): 374-392.

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, 25(4): 359-93.

Sunstein, C. (2013). Deciding by default. *University of Pennsylvania Law Review*, 162(1): 1-57.

Sunstein, C. (2014). *Conspiracy theories and other dangerous ideas*. Simon and Schuster.

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, *17*, 202–227.

Swami V., Chamorro-Premuzic T., & Furnham A. (2010). Unanswered questions: A preliminary investigation of personality and individual difference predictors of 9/11 conspiracist beliefs. *Applied Cognitive Psychology*, 24(6): 749–761.

Thomas S. B., & Quinn S. C. (1991). The Tuskegee syphilis study, 1932 to 1972: Implications for HIV education and AIDS risk education programs in the black community. *American journal of public health,* 81(11): 1498–1505.

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to Future Worth Wanting*. Oxford University Press.

Wheeler, C., Green, M.C. & Brock, T.C. (1999). Fictional Narratives Change Beliefs: Replications of Prentice, Gerrig, and Bailis (1997) with Mixed Corroboration. *Psychonomic Bulletin & Review* 6: 136–141.

Weinmann, M., Schneider, C., and vom Brocke, J. (2016). Digital nudging. *Business & Information Systems Engineering*, 58(6), 433-436.

Whitson J. A. & Galinsky A. D. (2008). Lacking control increases illusory pattern perception. *Science*. 322(5898): 115–117.

Zajonc, R. (1968). Attitudinal Effects Of Mere Exposure. *Journal of Personality and Social Psychology*. 9(2): 1–27.