Master's thesis
Master's Programme in Data Science

Auditing TikTok's Recommender System with Sock Puppets

Onni Aarne

October 16, 2022

Supervisor(s): Dr. Matti Nelimarkka

Examiner(s): Prof. Hannu Toivonen

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Onni Aarne

Työn nimi — Arbetets titel — Title

Auditing TikTok's Recommender System with Sock Puppets

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	October 16, 2022	38

Tiivistelmä — Referat — Abstract

The content we see is increasingly determined by ever more advanced recommender systems, and popular social media platform TikTok represents the forefront of this development (See Chapter 1). There has been much speculation about the workings of these recommender systems, but precious little systematic, controlled study (See Chapter 2). To improve our understanding of these systems, I developed sock puppet bots that consume content on TikTok as a normal user would (See Chapter 3). This allowed me to run controlled experiments to see how the TikTok recommender system would respond to sock puppets exhibiting different behaviors and preferences in a Finnish context, and how this would differ from the results obtained by earlier investigations (See Chapter 4). This research was done as part of a journalistic investigation in collaboration with Long Play. I found that TikTok appears to have adjusted their recommender system to personalize content seen by users to a much lesser degree, likely in response to a previous investigation by the WSJ. However, I came to the conclusion that, **while sock puppet audits can be useful, they are not a sufficiently scalable solution to algorithm governance, and other types of audits with more internal access are needed** (See Chapter 5).

ACM Computing Classification System (CCS):
Social and professional topics → Professional topics → Management of computing and information systems → System management → Technology audits

Avainsanat — Nyckelord — Keywords

TikTok, algorithm auditing, algorithmic accountability, recommender systems

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Acknowledgments

This work was generously supported by funding from the Helsingin Sanomat Foundation. I also wish to acknowledge CSC – IT Center for Science, Finland, for providing computational resources.

Contents

1 Introduction **1**

2 Related Work **5**

2.1 WSJ TikTok investigation 5

2.2 Algorithm audits 6

2.2.1 Recommender systems 7

2.2.2 Sock puppets 8

2.3 TikTok 8

3 Methodology **11**

3.1 System implementation 11

3.1.1 Technology stack 11

3.1.2 Avoiding detection 12

3.1.3 Technical difficulties 13

3.2 Experiment design and data collection 13

3.3 Data Analysis 15

3.4 Limitations 16

4 Results **19**

4.1 Personalization 19

4.1.1 Political rabbit holes 22

4.1.2 WSJ replication 23

4.2 Differently expressed preferences 24

5 Conclusions **27**

5.1 Sock puppet audits can be useful 27

5.2 Sock puppet audits cannot scale 28

5.3 Unauthorized audits have important limitations 28

5.4 Audits with better access are necessary 30

5.5 Audits need solutions 30

Bibliography **33**

Appendix A Sock puppet configuration details **37**

1. Introduction

In recent years, we have seen a growing flurry of public discussion and academic research on concerns related to the growing role of algorithmic systems in our lives. One strain attempting to bring rigor to this discussion have been unauthorized external audits of various public facing algorithmic systems, such as search engines [1] and video recommendation systems [2, 3, 4]. (Note that by unauthorized audits I mean audits conducted without the cooperation of the auditing target, as opposed to traditional external audits where the auditors are given access to internal documents of the organization being audited.) Some of the most famous such audits has been conducted by journalists [5, 6, 7]. A relatively recent example is a Wall Street Journal investigation of TikTok and its recommender system, which used "sock puppets", meaning accounts controlled by a computer program but which emulate the behavior of ordinary users. The investigators collected data about which content was recommended to different sock puppets exhibiting different interests [8]. This data showed strong "rabbit hole" effects, meaning that sock puppets which exhibited a preference for a specific type of content were soon shown almost exclusively that type of content [8]. (See Section 2.1 for a more detailed description.)

While this work has been invaluable for drawing attention to various issues related to algorithmic systems, and at times even motivating the operators of these systems to make changes, it has also been plagued by problems related to transparency and replicability, as well as rigor. Especially the journalistic work has not always been replicable. For example, the WSJ investigation [8] mentioned above did not publish a detailed description of which exact experiments they ran, and what their exact results were. This makes it more difficult for other to attempt replications of the experiments, or re-analyzing the data to scrutinize the WSJ's conclusions. Both academic and journalistic work has at times failed to appropriately account for the limitations of the approaches used. For example, causal language is often used, despite the methods used being unable to establish the causal role of the algorithm, and the external validity of the work is often overstated.

More generally, the field has failed to reflect on the effect it is having. Insofar as the goal of algorithmic accountability reporting and research is to actually hold

these systems and their operators to account and thus push these systems to become more beneficial to society, we must consider how effective different approaches are for achieving this goal. Additionally, it is important to develop a clearer understanding of the core problems related to algorithmic systems, so that accountability efforts can focus on the most significant problems, and call for changes that would meaningfully address those problems. The resources devoted to auditing are currently very limited, especially compared to the resources devoted to developing these systems, and their massive impact on society. This makes it especially valuable to direct those resources most effectively.

Thus there is a need for academic replicability and rigor to complement the wider reach and societal engagement of journalistic work. There is also a need for more thoughtful assessment of the epistemic advantages and limitations of different types of auditing approaches. More generally, it is also necessary to take stock of the limits of unauthorized external audits, and assess what other work is needed to form an effective portfolio of governance approaches for managing these ever-more influential systems.

To this end, this thesis presents a description of the methods and results of a journalistic sock puppet investigation of TikTok in a Finnish context, inspired by the Wall Street Journal's investigation [8]. As a check on the existing public discussion and academic literature, the investigation attempts to replicate the core findings of the WSJ investigation, as well as testing some common beliefs about problematic behaviors of recommender systems, such as intense "bubbles" [8, 9] and radicalizing "rabbit holes" of political content [2, 10] (see RQ 1a, Section 4.1). To demonstrate how auditing could be better informed by hypotheses based on specific theories of what could be going wrong with machine learning systems, I also performed some original experiments in an attempt to audit TikTok's recommender system for a particular type of problematic behavior (see RQ 2 and Section 4.2).

To conduct the investigation, so called sock puppets, or bot controlled accounts, were created to browse content on TikTok and collect data about the content they were shown. Different sock puppets were programmed to behave differently, allowing experiments to be run to determine how the system would respond to different user behaviors. The investigation was conducted in collaboration with Long Play, a Finnish organization specializing in longer, investigative journalistic pieces, and resulted in a journalistic article [11].

More precisely, my investigation addressed the following research questions:

1. How intensive is TikTok's personalization? (Section 4.1)

   (a) How quickly would TikTok learn the sock puppet's preferences, and how

much of the content seen by the sock puppets would relate to their interest once the preference had been learned.

(b) Would a sock puppet interested in politics be sent down a "rabbit hole" of such content? (Section 4.1.1)

(c) Could I replicate the key results of the WSJ investigation? (Section 4.1.2)

2. How does TikTok respond to different signals of interest? (Section 4.2)

(a) Does it respond faster to likes or watch time?

(b) Are likes or watch time weighed more heavily in determining which types of content to show a user?

RQ 2, about how TikTok responds to different types of information about a user's preferences, is particularly interesting due to its implications for user agency and welfare. More thoroughly considered decisions can be seen as better reflecting a person's idealized preferences, the preferences which they would endorse upon reflection. The decision to buy a book is usually made with at least some minutes of thought; Before following someone on Twitter, we might vet their latest Tweets to check whether we think their Tweets are worth including in our feeds. On TikTok, however, it appears that what we see in our feeds is largely determined by the split second not-even-a-decision of which of the algorithm-curated videos we linger on. This suggests that TikTok's recommendations may reflect our idealized preferences especially poorly. This also suggests that TikTok may make it particularly difficult for users to deliberately influence the content of their feed.* For this reason I wanted to test how TikTok weighed information gleaned from decisions of how long to spend watching a video, as opposed to the more deliberate and considered choice of liking a video. While this may seem like a minor concern in the case of TikTok, as more and more of what humans see and do are influenced by increasingly sophisticated algorithms, subtle erosions of human agency and drifts away from what we would truly want can compound to have a significant influence on the future [12].

The investigation found that the sock puppets experienced little to no personalization, and certainly were not led down rabbit holes, as was observed in the Wall Street Journal investigation. This non-effect was consistently observed across many

*Of course reality is a bit more complicated than this. Other factors, such as the amount of relevant information one has access to, can also affect the quality of one's decisions. One might also argue that collecting information about a larger number of faster decisions gives the system more data about the user's preferences and might thus allow the system to infer a better, if less idealized, representation of those preferences. Nonetheless, the question of which decisions a recommender system bases its inferences on remains an interesting one.

different behavioral profiles, expressing many different interests. It is most likely that this difference in results were driven by changes made to the TikTok recommender system.

Beyond the object-level questions it addresses, this investigation is also used as a starting point to critically examine the advantages and limitations of this type of audit. This is done by assessing the epistemic advantages and limitations of these methods, the technical difficulties involved in sock puppet investigations, as well as whether past investigations have caused TikTok to change.

The structure of this thesis is as follows: Chapter 2 describes previous work and motivating background for this thesis, including a more detailed description of the WSJ investigation that inspired our investigation (Section 2.1). Chapter 3 describes my sock puppet system (Section 3.1), its experimental capabilities (Section 3.2) as well as limitations (Section 3.4). Chapter 4 describes the experiments making up the investigation, and their results. Finally, Chapter 5 seeks to leverage experience gained from this investigation to answer questions about the overall feasibility of this approach to auditing as a solution to the problems of accountability.

2. Related Work

In this chapter, I will begin by discussing the Wall Street Journal's TikTok investigation [8], which heavily inspired my work. I will then cover previous work on algorithm auditing in Section 2.2, particularly what other methods have been used in previous external audits of recommender systems, and previous auditing work applying sock puppets. Additionally, previous work on TikTok and its impact will be discussed in Section 2.3.

2.1 WSJ TikTok investigation

The Wall Street Journal (WSJ) conducted an investigation of TikTok's recommender system, published under the title *How TikTok's Algorithm Figures Out Your Deepest Desires* in July of 2021 [13]. The investigators created a number of sock puppet accounts on TikTok that watched videos on their for you page (FYP) as a normal user would. These sock puppets were programmed with different interests, and would preferentially watch videos that had tags related to their interests.

The investigation found that most of these sock puppets were pushed very quickly deep down "rabbit holes" of content related to their particular interest. TikTok's recommender system started showing these sock puppets nearly exclusively content related to the particular interest the sock puppet was programmed to have after just a few hours of watch time. In particular, the investigation found that sock puppets would be shown videos with fewer and fewer views over time as TikTok's recommender system learned the sock puppet's preference and served them increasingly niche content [13]. The WSJ also published a number of other articles describing particular findings from the same investigation in more detail [8, 14].

The published video describing the investigation describes in the most detail the trajectory of a particular sock puppet interested in depression related content. The sock puppet, apparently a 24-year-old from Kentucky, would watch videos tagged '#depression', '#anxiety', '#sad', '#pain' or similar twice, and other videos once. After only 224 videos and 36 minutes of this sock puppet scrolling through TikTok, 93% of the videos it was shown were depression related [13]. This is a chilling example

of the potential harms from "rabbit holes", but it should be kept in mind that this sock puppet exhibited an interest exclusively in depressive content, which is unlikely to be the case for many human users. In Section 4.1.2 I describe my attempt to replicate this particular experiment as closely as possible, and find very different results than they did.

In December 2021, TikTok announced that they are adjusting the recommender system to "diversify recommendations", and "testing ways to avoid recommending a series of similar content – such as around extreme dieting or fitness, sadness, or breakups – to protect against viewing too much of a content category that may be fine as a single video but problematic if viewed in clusters" [15]. The timing and the particular topics mentioned strongly suggest that this was in response to WSJ's reporting, but this is not directly acknowledged by the TikTok announcement.

2.2 Algorithm audits

This subsection draws heavily on a 2021 systematic literature review of the algorithm auditing literature up to 2020 by Bandy [16]. Specifically, Bandy covers academic articles describing studies that look for problematic behavior in deployed algorithmic systems. These studies are typically unauthorized, meaning that they were conducted without collaboration from the audited party, which is typically a large internet company, such as Google or Meta.

Bandy notes numerous previous audits that focused on recommender systems or used sock puppets, but notes that despite its growing popularity, he could not find any audit of TikTok [16].

Bandy identifies four types of problematic behavior that previous audits have focused on: discrimination, distortion, exploitation, and misjudgment. Of these, the issues that my investigation focuses on, and that other have raised about TikTok would most naturally fall under what Bandy calls distortion: hyper-personalization and misleading content. But a primary concern of my investigation is absent from Bandy's typology: whether the algorithm leads to effects contrary to the user's considered self-interest, or contrary to their flourishing. This appears to be driven largely by the fact that audits typically look for "problematic behavior": Behavior that clearly violates some ethical norm, such as fairness or privacy. Systems which are only mildly or diffusely bad for their users are generally not considered problematic, despite the fact that they may, over time and in the aggregate, have significant detrimental effects.

2.2.1 Recommender systems

Bandy's review [16] found eight audits of recommendation algorithms. Bandy reports that audits of recommender systems have primarily focused on echo chamber effects and source concentration: the tendency to primarily recommend content from a small number of sources. Bandy found that most audits looking for echo chamber effects had failed to find significant effects [16]. In this context, the findings of the WSJ investigation are notable for having found an extremely strong and rapid echo chamber effect [13].

The studies found by Bandy et al. that are most similar to my investigation are audits of YouTube, particularly its recommender system [2, 3]. I will also consider some more recent papers not included in their review [4, 17].

No study of YouTube has, to my knowledge, used true logged in sock puppets. Some studies, such as Ribeiro et al. [2] scrape recommendations from the YouTube website without being logged in. Many papers, such as Papadamou et al. [17] use the YouTube Data API's ability to query for videos related to a given video as an approximation of recommendations. These related videos are likely based on some notion of similarity between videos which acts as an input to YouTube's recommender system, but which is not the system itself: It does not implement any kind of personalization, and has only a loose relationship to what even a logged-out real user would see as recommendations. This obviously gives much weaker evidence about the real-world impact of the recommender system than scraping or sock puppets.

In addition to their investigation of recommendations, Ribeiro et al. [2] also used an approach related to crowdsourcing. They scraped comments from videos, and analyzed how commenters moved between channels, finding a phenomenon of radicalization. This type of data is valuable in providing strong evidence of a phenomenon of radicalization, but cannot tell us about the causal role played by the recommender system. Of course, even this type of data has its limitations: Only a small minority of users comment on videos on YouTube, and it does not seem implausible that those who comment would systematically differ from those who do in a way relevant to the findings of Ribeiro et al.

Another approach to studying recommender systems would be to crowdsource data from real users. This approach has the advantage of giving more realistic data, but usually the data cannot be used to establish causation. A notable example of crowdsourcing is The Markup's Citizen Browser project [18]. The project collects data about what a representative panel of US citizens see in their Facebook feeds, and The Markup has written about how this results in a polarized news landscape [19], and recommends types of comment Facebook claimed not to recommend [20].

2.2.2 Sock puppets

A systematic literature review of algorithmic audits by Bandy [16] finds 12 studies using sock puppets published up to 2020.

Compared to scraping, sock puppets can capture personalization, and compared to crowdsourcing, sock puppets enable the establishment of causal relationships between aspects of user behavior. For example, sock puppets have been used to show how the ads shown to a user are affected by the user's reported race or gender, all else kept equal [21].

Sock puppets still have their limitations. If the behavior of sock puppets is sufficiently different from any real humans, the findings of suck puppet investigations might not generalize. This is especially concerning when investigating systems based on highly nonlinear types of machine learning systems, such as deep learning systems.

Ultimately the greatest limitation of sock puppet studies is how difficult they are to execute in practice. One aspect of this is the basic technical difficulty of writing bots or sock puppets compared to using an authorized API or even an unauthorized scraper. The second aspect is that most platforms are actively hostile to bots and sock puppets. Sections 3.1 discusses some of these challenges related to the technical implementation of sock puppet investigations in the context of my own investigation.

2.3 TikTok

Relative to the popularity of TikTok, there is a rather limited amount of rigorous academic research on it. I will highlight some research most relevant to my own investigation.

Bandy and Diakopoulos [22] investigate the impact of TikTok's recommender system on collective political action in the case of a particular viral video. They find that the algorithm does not systematically amplify content with such calls to political action, but does enable them to reach very large numbers of people very quickly.

Klug et al. [23] interviewed U.S. TikTok users to understand their folk beliefs about the recommender system, specifically regarding which types of content the system would promote to a large number of people. They found that users believed that video success was influenced by video engagement, posting time, and adding trending hashtags. Klug et al. [23] additionally attempted to test these user beliefs by scraping popular videos and looking at relationships between engagement metrics and view count in that sample. However, because they select on the dependent variable (video popularity), it is dubious whether the correlations they found would generalize beyond that biased sample. Even if we are willing to believe that they do generalize, correlation

need not imply causation.

To truly test such folk beliefs, one would need to run controlled experiments. In the case of factors affecting video success, this would mean randomizing those factors on videos before posting them. Sock puppets allow such controlled tests of which aspects of user behavior affect what content they see.

3. Methodology

In this chapter, I will describe the implementation of my sock puppet system, and its scalability in Section 3.1. Then in Section 3.2 I describe the system's capabilities in terms of behavior and data collection, and the types of experiments the system can be used to conduct. Finally, in Section 3.4 I discuss the limitations of the system.

3.1 System implementation

My sock puppet system was implemented using browser automation, and designed to mimic normal human users. This type of approach is necessary to avoid being blocked by the site's bot detection systems, but it also made for a fragile system that required a significant amount of human labor to set up each experiment, thus greatly undermining the scalability of the system.

3.1.1 Technology stack

To implement my sock puppets, I used the browser automation tool Selenium, which allows a computer program to control an instance of a web browser. This allows my system to read the content of the TikTok website and interact with it through clicking on certain site elements or entering certain keystrokes. In practice, the primary functions of the system are to read some key information about a video from certain elements of the page, and then press a key to move to the next video after a time. (The details of the sock puppet's behavior are discussed in Section 3.2.)

Thus, from the point of view of TikTok, the sock puppet appeared to be a person browsing TikTok using the Google Chrome browser on a desktop computer running Ubuntu Linux somewhere in central Finland. The browsing behavior of the sock puppets is further discussed in Section 3.2, while the following subsections provide further detail on the technical implementation of the system, and associated complications.

On a technical level, my system consists of a Python program that runs on a server in the CSC Pouta cloud, as a headful Selenium instance using `undetected_chromedriver` using a VNC server to instantiate a virtual desktop on

the server. The code for the system is available to researchers on request.

The system operates primarily based on information about e.g. video length and video description. This is primarily collected directly from the page's HTML, but some metadata, such as exact view and share counts, is not possible to directly scrape from the user-facing UI and is retrieved in real time using TikTokApi. The system then logs information about which videos the sock puppet saw at which time, how they behaved in response to each video, and what the metadata of the videos were at the time.

3.1.2 Avoiding detection

The system needed to be set up in a particular way to avoid an automated browser being detected as such. The system runs on a virtual desktop, and uses a headful Selenium instance. Headfulness means that an actual graphical browser is fully rendered, as opposed to a headless browser, which would be a simulated browser that does not render an actual graphical interface. Headless operation requires less computing resources, and is typically preferred for browser automation, but as a headless browser could be detected as such by the websites visited, and blocked. Some features of complex dynamic websites like TikTok might also not work on a headless browser. Additionally, a special variant of the official Selenium driver for Chrome, called `undetected_chromedriver` was used to control the browser instance, because the official driver can be detected by websites.

The behavior of the sock puppets as they used the site also needed to be specifically designed to be human-like. For example, the sock puppet primarily navigates by clicking on buttons on the page, rather than simply opening the desired URL directly.

The account creation and logging in steps must be conducted by a human, because this step requires complex interactions which would be difficult to program a sock puppet to do naturalistically, including the completion of a simple CAPTCHA. When the platform has thus been reassured that the account belongs to a real human, the main work of watching videos can be delegated to a simple sock puppet.

Keyboard shortcuts can be used to control the basic action loop of watching a video, optionally liking it, and moving to the next video. Human and machine key presses are already difficult to distinguish, and Gaussian noise was added to the timings of they key presses to make them even more naturalistic.

Each sock puppet additionally had its own set of cookies, and each account was made with a different email address.

3.1.3 Technical difficulties

One downside of using sock puppets like this, compared to e.g. collecting data through an API, is that browser automation is comparatively unreliable. Most core functions of the system had to be implemented twice or more in different ways, in case the primary method failed, seemingly randomly. For example the `pyautogui` library was often used to for fallback actions by moving the actual cursor to particular manually specified coordinates in cases where clicking on elements selected by Selenium failed.

The most natural way to implement browser automation is to refer to page elements using CSS selectors and other identifying factors from the page source. A problem with this is that many modern websites are also actively hostile, and dynamically generate class identifiers to hinder bots and sock puppets. This has in some cases bled over to sabotaging accessibility features [24].

As mentioned in the previous subsection, the system has to run headfully, which increases both the computational cost and complexity of the system, as a virtual desktop is required. Selenium also appears to be substantially less stable when running headfully.

One technical limitation of my system that could have affected my results is that all of the sock puppet users appeared to be using systems very similar to each other, and there is some chance that this led TikTok to assume that they were all one user. The puppets used a small number of different IP addresses, often with many puppets running from the same IP simultaneously. Further, their "fingerprint", meaning the details of which operating system and browser they were using, which browser extensions they had installed, etc. were all very similar or identical to each other. They all also had similarly strange cookie profiles in that they had no cookies for any sites other than TikTok (though each had distinct TikTok cookies.) I do not believe that this is significant enough that it could have completely invalidated my results, but it may have added some amount of noise or distortion.

3.2 Experiment design and data collection

The primary way to browse TikTok is through the "for you" page, commonly known as the FYP. This is what my sock puppets do. The FYP consists of a sequence of videos selected for the user by TikTok's recommender system. As the user swipes to the next video it automatically starts and will keep looping until the user swipes to the next video. The recommender system then observes the user's reactions to the videos to determine what type of content the user is interested in, and will show more of that type of content.

My sock puppets would decide how to respond to each video based on processing the video description and other metadata. Each sock puppet* had one or more interests, defined by a list of tags and keywords, and it would respond in different ways to videos if they matched it's interests, i.e. if one or more of the tags or keywords was found in the video description.

This allowed me to test how differently expressed interests affected the content served to the sock puppets. I designed specific behavior profiles to address my research questions. For example, the basic "Lingerer" behavior profile simply watches videos that match its interest twice, and skips† others. It can be used to test how quickly the recommender system will discover a given interest based on this behavior and how much of the content served to the sock puppet will eventually match that interest (RQ 1a. The different behavior profiles I used are described in more detail in Table 3.1.

My research question 2, about how the recommender system responds do different types of behavioral signals of interest, can be tested by e.g. comparing a "Lingerer" sock puppet to a "Liker" with the same interest. The "Conflicted" behavior profile is designed to test this question more directly by giving one sock puppet two interests that are expressed by different signals (See Section 4.2). The "Baseline" behavior profile gives another a point of comparison for how much of different types of content a sock puppet with no particular interests would see. The specific experiments I conducted and the results I got are described in Chapter 4.

To help the recommender system get going, some experiments were additionally primed with a tag: They would begin their life by watching n, typically 20, of the first videos that TikTok returned at the time when one searched for content matching the tag.

The system could also be easily expanded to allow additional behavior profiles, e.g. a profile that only skips videos that it is specifically disinterested in, or one that exhibits a preference for videos of a particular length. Further development of the system could also introduce an additional behavior of pausing viewing and clicking away from the window entirely, if, for example, the sock puppet had recently seen too many videos of a type it disliked, or too few of a type it was interested in.

*Except for the baseline sock puppets.

†If a sock puppet chose to skip a video as uninteresting, this did not actually mean moving on as fast as possible, because this could be unrealistically fast. Instead, uninteresting videos were watched for a varying fraction of the video's length, at most half the video length or 15 seconds, whichever is lower.

Behavior Profile	Description
Baseline	Watches every video once
Lingerer	Skips all videos except those matching its interest, which it watches twice.
Liker	Skips all videos except those matching its interest, which it likes and watches once.
Conflicted	Skips all videos except those matching one of its two interests. Anything matching the first interest is liked and watched once, anything matching the second interest is watched twice.
WSJ Lingerer	Watches all videos once, except those matching its interest, which it watches twice. This is the behavior used by in the WSJ investigation.

Table 3.1: Detailed descriptions of behavior profiles.

3.3 Data Analysis

To determine how much the recommender system adapted the content of the feed to match the interests of a sock puppets, one can analyze how the amount of interest-matching content changed over the lifespan of the sock puppet, and how it compared to a baseline sock puppet that watched all videos equally patiently. See Chapter 4 for how I did this in my investigation.

Due to the small sample size, my data analysis approach was mostly qualitative. As can be seen in the figures throughout Chapter 4, I plotted the rolling sum of the number of videos out of the last n videos that matched a given interest. When drawing such a plot for the interest that e.g. a Lingerer sock puppet had, one would expect an initial upward trend stabilizing at a level clearly above the curve that would be drawn for the same interest and a baseline sock puppet.

More sophisticated analysis would have been valuable if I had been trying to discern subtle variations in the data, but in this case the non-effect is quite clear, compared to the very strong effect that one could have expected, and that was reported by the WSJ investigation.

3.4 Limitations

Sock puppet audits in general, and my approach in particular, suffer from numerous limitations. In addition to the technical difficulty of completing an audit like this, which I discussed above, there are some limitations inherent to my system and to sock puppet audits like this in general. While these limitations are significant, I do not believe they significantly undermine my core findings.

A basic limitation of my approach was that the sock puppets were restricted to interests that were reliably expressed in a video's description, usually in its tags. Many common types of content on TikTok, such as viral dances and comedic content, are not usually tagged with anything that would directly indicate the type of the video.* Fortunately there are some subtypes of videos, such as the types I used as interests for my sock puppets (see Table A.1), that are consistently indicated by one of a few genre-indicating tags, such as #cleantok or #tiktokfood. However, this means that experiments like this are restricted to emulating users who are interested only in videos belonging to such well-defined and self-aware subgenres, which represents only a minority of all content on TikTok

A key limitation of one-off studies like this is that systems like the TikTok recommender are continuously changing. TikTok's engineers are constantly implementing tweaks and improvements. A/B tests, where some users are randomly exposed to one version of a system and others another, mean that even a set of sock puppets run at the same time may be not all be interacting with the same version of the system. Production machine learning systems are often continuously or intermittently re-trained with fresh data about recent user behavior, so the system would be changing even without active engineer intervention.

The behavior of the algorithm also changes over space, in addition to time. Likely the most significant limitation of the usefulness and generalizability of these results is that almost all of my experiments were conducted in Finland. It is likely that the vast majority of the engineering effort that has gone into developing the recommender system has been focused on improving the experience of users in the U.S. and China. The Finnish context may confuse the algorithm in some ways. The amount of Finnish content it has to pick from is much more limited than in the anglosphere. For example, the total number of people actively posting about Finnish politics on TikTok may be as low as a few dozen, though this is very difficult to estimate. There is a real

*Tags, in general, are a somewhat strange phenomenon on TikTok. While tags were originally intended to help users find content by browsing by tag, on TikTok tags seem to have morphed almost entirely into a way for creators to attempt to influence whether the recommender system will show their video to users, and to which types of users [23].

possibility that this affected some of my experiments. The system is also operating in a more challenging environment in that international English-language content has significant appeal in Finland, and thus the system is dealing with a more complex, bilingual environment compared to the anglosphere.

A key advantage of sock puppets is that they "experience" the system as a real user would. However, for the puppet's experience to be realistic, *id est* representative of actual user experience, its own behavior must be realistic as well. This is easier said than done in the absence of detailed data about real user behavior. For example, as discussed in Section 3.2, one design question I faced in defining my sock puppets' behavior was how long to linger on videos deemed uninteresting. Skipping to the next video as soon as my system was able to read the tags of the video would be unrealistically fast. At the other extreme, the sock puppet could watch all videos once by default, even if not particularly interested. For example, my "baseline" behavior profile watches each video exactly once. But this likely makes my baseline appear to have a preference for longer videos, compared to a real users, who likely have finite patience. However, I believe these biases are sufficiently subtle as not to have significantly affected my results.

A more concerning consideration is the length of viewing sessions. My sock puppets generally watched TikTok in marathon sessions, for hours at a time. Real users, on the other hand, probably often watch in shorter sittings of perhaps a few minutes. This might importantly affect the behavior of the recommender system. For example, the fact that my sock puppet keeps browsing, instead of logging off, could be interpreted by the recommender system as a sign that the user is happy with the content it is being shown.

One possible explanation of the lack of personalization I observed in my experiments is that the recommender system may not allow more than some fixed number of videos on the same topic, or from the same creator, per session. Indeed, based on my personal experience using TikTok, the recommender seems to fastidiously avoid showing two videos from the same creator in a row. This suggests that a sock puppet that watched in shorter sessions, perhaps distributed over a greater number of days, might have had a meaningfully different overall experience. For similar reasons, the experiences of sock puppets with only one or two interests may not generalize to those of actual users who likely have numerous interests. I did not have the time to experimentally test these possibilities, and they are likely only a few among many ways in which my sock puppet's acted in very unusual ways.

A further concern related to this comes from the fact that the TikTok recommender system is likely based on deep learning. Deep learning systems are known to be brittle, and often behave erratically on inputs that are well outside their training

distribution.Due to the nonlinear and non-monotonic nature of these systems, any regularities discovered across sock puppet experiments need not, in principle, generalize at all to the range of inputs generated by real users.

Some aspects of the experiments were not tightly controlled. For example, the experiments were often run at whatever time was convenient for the experimenter, including over night. It is possible that TikTok made some inferences about the sock puppet accounts based on this, but it seems very unlikely to me that this meaningfully affected the results. As mentioned above, the length of the viewing sessions of the sock puppets was often determined by a combination of whatever was convenient for the experimenter, and when my system crashed.

A very basic limitation of my results is also that I only used the web version of TikTok on a desktop browser. It is possible that the web experience is in some way different from the mobile in-app experience, and the in-app experience is likely the experience of the vast majority of users. For example, the app may be able to track user behavior with more detail and accuracy in a browser, or certain forms of behavioral tracking may simply not have been implemented in the browser, because most users are on the mobile app. However, it seems unlikely that a core feature like content recommendation would significantly differ between platforms. Additionally, TikTok may be using the information that my sock puppet is browsing on a Linux desktop to infer something about the nature of that user*.

Ultimately, while all of these considerations mean that the design of experiments like this are perhaps more art than science, I think it is likely that the basic findings of my experiments still hold: My sock puppets did not experience the kind of intense rabbit holes that have been observed before, and that would have been predicted by folk theories of the algorithm.

*Sadly I did not see any Linux-related TikTok content among the data collected by the sock puppets.

4. Results

This section describes a series of experiments and their results. Section 4.1 describes experiments intended to answer RQ 1 about the intensity of TikTok's personalization, while Section 4.2 covers experiments addressing RQ 2 about how the recommender system responds to different expressions of preferences.

The exact configurations of the experiments I ran are described in Table 4.1, and the results of the experiments are described in Table 4.2. The details of the experimental capabilities of my sock puppet system and the different behavior profiles, as well as my data analysis approach, were covered in the previous chapter, in Sections 3.2 and 3.3. Further details of the specific configurations used can be found in Appendix A. The data collected in the experiments is available upon request.

4.1 Personalization

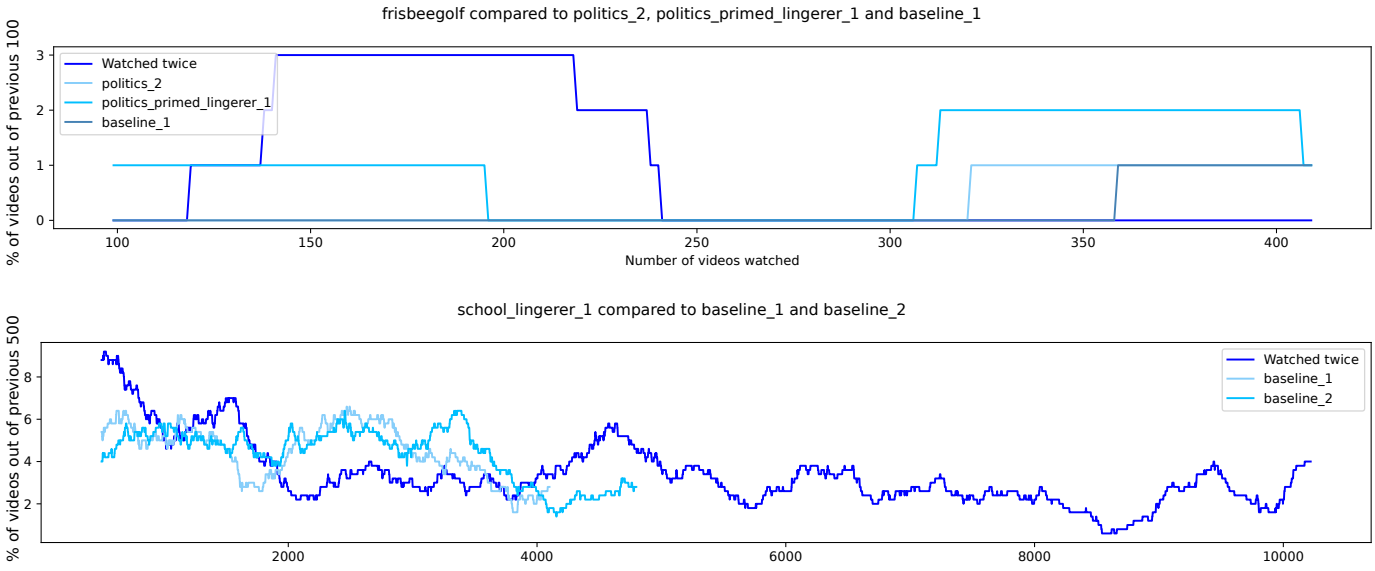

Figure 4.1: Plots of the density of relevant content seen by my sock puppets interested in frisbeegolf (top) and school (bottom) related content.

Experiment name	Behavior profile	Interests	Priming	Began	Ended
politics_1	Lingerer	Politics	N/A	10-31 22:32	11-01 09:04
politics_2	Lingerer	Politics	N/A	11-03 13:41	11-04 23:30
frisbeegolf	Lingerer	Frisbeegolf	N/A	11-03 17:04	11-03 18:28
politics_primed_lingerer_1	Lingerer	Politics	#politiikka	11-10 23:29	11-15 12:54
politics_primed_liker_1	Liker	Politics	#politiikka	11-14 00:00	11-16 19:04
politics_primed_basic_1	Baseline	N/A	#politiikka	11-14 19:26	11-14 21:48
school_lingerer_1	Lingerer	School	N/A	11-18 14:25	11-27 05:08
conflicted_food_cleaning_1	Conflicted	Cooking, Cleaning	N/A	11-20 21:29	11-20 21:49
conflicted_cleaning_food_1	Conflicted	Cleaning, Cooking	N/A	11-20 21:40	11-27 05:06
conflicted_food_cleaning_1	Conflicted	Cooking, Cleaning	N/A	11-22 21:59	11-27 05:09
baseline_1	Baseline	N/A	N/A	11-23 17:35	11-27 05:10
baseline_2	Baseline	N/A	N/A	11-23 17:48	11-27 05:11
politics_primed_basic_2	Baseline	N/A	#politiikka	11-23 18:03	11-27 05:14
wsj_depression_atlanta	WSJ Lingerer	Depression	N/A	11-26 19:38	11-27 03:46

Table 4.1: The behavioral profiles of each of the experiments, in chronological order by start time. In every experiment where priming was used, the number of videos watched at the priming step was 20. The meaning of the terms in the behavior profile column are explained in Table 3.1. Table 4.2 gives more details about results of the experiments.

As discussed in Section 2.1, the WSJ investigation found very intensive personalization. Their sock puppets were quickly sent down "rabbit holes" and shown increas-

ingly niche content, ever more specifically related to the specific interest of the sock puppet. (See Section 2.1 for a more detailed description.) Many of my experiments aimed to assess the speed and intensity of this personalization. I found the opposite: that my sock puppets experienced little to no personalization, across multiple different interests and behavior profiles. This was also true for my best attempt to replicate one of the WSJ investigation's experiments as faithfully as possible (Section 4.1.2).

As basic test of the recommender system, I created sock puppets interested in topics that I knew they were likely to see in their feed, based on my own observations of the platform, and that could be relatively reliably identified by their tags: Frisbee golf, and school*. Both of these sock puppets expressed their preference by watching videos twice if the video had a tag related to their interest, and skipping other videos. The deep blue line in the two plots in Figure 4.1 plot, over the lifespan of the sock puppet, the rolling sum out of the last n videos seen by the sock puppet over their lifespans that directly related to the particular interests of the sock puppet. For comparison, the lighter blue lines show the rolling sum of videos related to the relevant topic (frisbeegolf or school) seen by other sock puppets with no or different interests. For example, from the top plot in Figure 4.1 the frisbeegolf sock puppet (deep blue) saw a total of three videos related to its interest, which is equal to the number of frisbeegolf related videos seen by politics_primed_lingerer (sky blue) over the same span of time. This shows that the recommender system did not significantly respond to the preference exhibited by the sock puppet interested in frisbeegolf.

Overall, we can see that content related to their interests never dominated their video feeds, and indeed the overall amount of videos they saw that related to their interests was not obviously higher than that seen by other sock puppets with different interests. While it appears that the sock puppet interested in school-related content saw more school related content very early it its life than the baseline sock puppets it was compared to, with this few examples, we cannot reliably distinguish this from noise. Indeed, if we reduce the window over which the rolling sum is computed to 20 videos, we observe that while the density of school-related videos peaks earlier for that sock puppet than the baseline, the peak density of 5/20 videos being school related is also repeatedly reached by baseline_2. One possible explanation of this is that the recommender system detected the preference early on in the sock puppet's life, but would not allow the rolling sum of school-related videos to exceed 25%, and then soon reduced the frequency of those videos, either because there were no more recent school related videos to show, or in an attempt to maintain diversity in content shown over the course of a session. However, this hypothesis of a 25 % cap in the frequency of a

*The exact tags that were considered school or frisbeegolf related, as well as the methods used to construct these lists, are documented in Appendix A.

specific topic does not explain why other sock puppet with less common interests, such as those interested in politics or frisbeegolf, never got close to that.

Many of the other experiments discussed in more detail in following section, such as politics_lingerer and politics_liker depicted in Figure 4.2, and the "conflicted" sock puppets, depicted in Figure 4.3, also provide evidence on the intensity of personalization. Similarly to the experiment discussed above, these did not experience clearly higher frequencies of the types of content they were interested in, compared to the two baseline experiments.

Overall, it appears that either the recommender system failed to infer my sock puppets' preferences, or for some reason refused to adjust their feeds enough for the personalization to be obvious.

4.1.1 Political rabbit holes

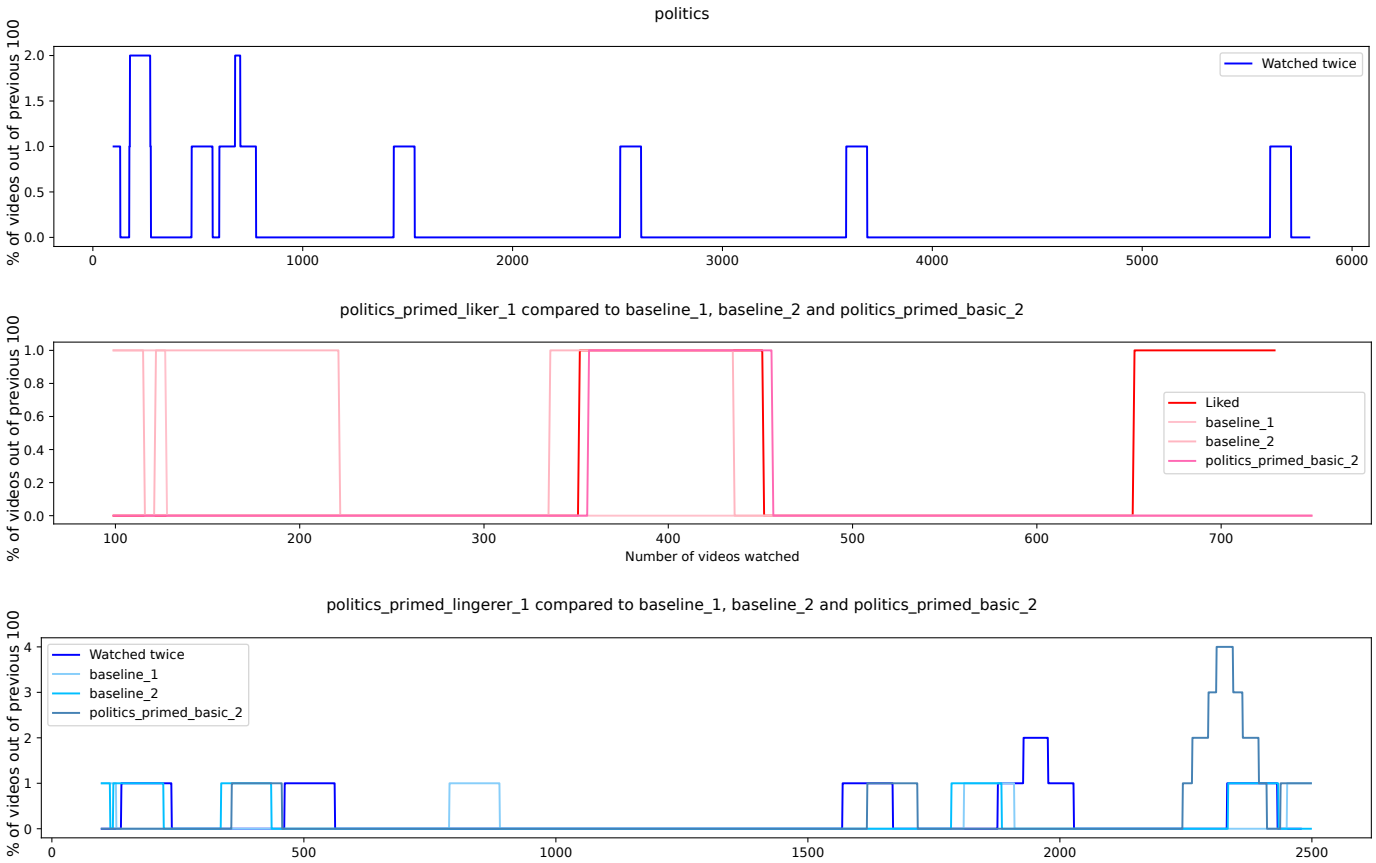

Figure 4.2: Plots of the density of politics-related content seen by the politics related experiments.

A personalization related question of particular interest was that of "rabbit holes" related to political content (RQ 1b). A secondary question under investigation, due to its journalistic interest, was also the nature of the political content that a hypothesized

rabbit hole would lead to. As discussed above, my experiments did not find any rabbit holes, political or otherwise, and thus the second question went unanswered.

I experimented with several sock puppets with interests in politics, whose lifespans are plotted in Figure 4.2. Each used the same set of tags* to detect political content†. This set of tags was made up of general Finnish words related to politics that I had observed to be in use on TikTok. This included words such as those referring to the political left and right‡, to the government§, and to all of the major parties. My first sock puppet interested in politics, "politics", expressed its interest through the Lingerer profile (See Table 3.1). Because this sock puppet did not see much political content, later experiments were primed by specifically searching out politics related videos before starting to browse their FYP ("For you" page, see Section 3.2), in an attempt to help the recommender system to get a clue of the sock puppet's interests. These later experiments included one using the Lingerer profile, one using the Liker profile, and one (politics_primed_basic) that underwent the priming procedure but then implemented the baseline behavior profile, meaning did not express any specific interests after the priming. The purpose of this last experiment was to measure the effect of the priming independently of later behavior by the sock puppet. As can be seen from Figure 4.2, none of these experiments saw clearly more politics related content than the baseline experiments, or experiments with different interests.

4.1.2 WSJ replication

One of my experiments was intended as as exact a replication of the headlining experiment from the WSJ TikTok investigation [8] (Section 2.1) as was possible within constraints I faced. The goal of the experiment was to answer my investigation's research questions 2a and 1c about TikTok's personalization, and to thus better understand whether publishing investigations like this can have an effect on the systems being investigated. To summarize, my experiment did not find the kind of rabbit hole effect that was observed in the WSJ investigation, and also observed far less depression related content. This strongly suggests that TikTok had indeed made changes in response to the publication of the WSJ investigation.

The sock puppet (wsj_depression_atlanta) was run through a popular consumer VPN in Atlanta, as this was the closest to the original Kentucky out of the locations

*These are listed in Appendix A

†Which I had, in my explorations, found to often indeed have politics-related tags.

‡The words for the political left and right (vasemmisto and oikeisto, respectively) are distinct from the words for the broader concepts of left and right.

§The Finnish word "hallitus" specifically refers to the government or administration set up by the governing coalition, rather than the government writ large.

available. At the account creation stage, verification with only an email address was not accepted as usual, and TikTok demanded a U.S. phone number. We did not have access to that, and thus resorted to conducting the experiment without the account being logged in. However, cookies were saved as usual, so it should have been possible for TikTok to track the session as normal. This sock puppet only completed one watching session.

During that session it saw only two depression related videos, and despite watching those twice as the WSJ sock puppet had done, the recommender system was not offered more of the same type of content (See Table 4.2). The first of these videos was seen around the 700th video. The WSJ sock puppet was deep in a rabbit hole (90% depression related videos) after only 200 videos. Indeed, my Finland-based baseline sock puppets saw many more depression related videos (6 and 4) than this replication attempt did, across the same number of videos watched*. This was of course just one run, but it strongly suggests that new users are shown less depressive content, and are not allowed down rabbit holes, compared to what the WSJ investigation observed.

4.2 Differently expressed preferences

A key research question in my investigation was whether different behavioral signals of preferences would be treated differently by the recommender system (RQ 2). As discussed in Chapter 1, my concern was that TikTok's recommender system could be misaligned with the considered preferences of users by disregarding or underweighting intentional and conscious feedback, namely likes, in favor of less conscious expressions of preferences, that the user would be less likely to endorse on reflection. TikTok's recommender system "passed" this audit on a technicality: It did not underweight interests expressed via likes, because it did not appear to "weight" any interests at all.

More specifically, the primed "liker" and "lingerer" politics experiments, shown in Figure 4.2, were intended to measure whether liking interesting videos and watching them once would lead to the recommender system "learning" that interest faster, or showing more videos related to the interest (RQ 2a). As discussed above, the recommender system never clearly learned these interests, so this question was left largely unanswered.

Two "conflicted" sock puppets, depicted in Figure 4.3 were intended to answer the question of how the recommendation system would weigh or prioritize two interests that are revealed by different behavioral signals. As detailed in Chapter 1, the motivation

*This might suggest that depression related content was specifically suppressed in the US in response to the WSJ investigation, but it could simply be due to the Finns' well known depressive tendencies.

Experiment name	Runs	Total hours watched	% liked	% watched once	% watched twice	% skipped
politics_1	4	N/A	0.00	0.00	0.00	0.00
politics_2	3	13.28	0.00	0.00	0.17	99.83
frisbeegolf	1	0.92	0.00	0.00	0.73	99.27
politics_primed_lingerer_1	9	5.46	0.00	0.80	0.24	98.96
politics_primed_liker_1	6	3.45	0.27	100.00	0.00	0.00
politics_primed_basic_1	1	2.01	0.00	100.00	0.00	0.00
school_lingerer_1	11	27.32	0.00	0.00	3.49	96.51
conflicted_food_cleaning_1	1	0.30	3.45	3.45	0.00	96.55
conflicted_cleaning_food_1	10	34.12	0.55	0.55	1.39	98.06
conflicted_food_cleaning_1	6	19.40	1.39	1.39	0.64	97.97
baseline_1	4	32.86	0.00	100.00	0.00	0.00
baseline_2	4	30.73	0.00	100.00	0.00	0.00
politics_primed_basic_2	3	25.39	0.00	100.00	0.00	0.00
wsj_depression_atlanta	1	5.84	0.00	99.82	0.18	0.00

Table 4.2: Details of the results of each experiment. The experimental configurations are described rin Table 4.1

for this question is that some more deliberate signals, such as likes, could be expected to better match what a person's more considered or higher-order desires are, as compared to less deliberate signals, such as watch time.

The "conflicted" sock puppets had two interests, cooking related content and cleaning related content. These interests were picked because they are relatively common on TikTok, easy to distinguish via tags, and relatively similar to each other in their frequency. Each expressed one interest through watching the video twice, and the other interest through watching the video once but liking it. Which interest was which was flipped between the experiments, to account for the one of the interests simply being more common or easier for the recommender system to learn, regardless of signal type. Uninteresting videos were skipped. As with the other experiments, the sock puppets did not see clearly more of the types of content they were interested in than baseline sock puppets, so no meaningful answer to the research question at hand can be given.

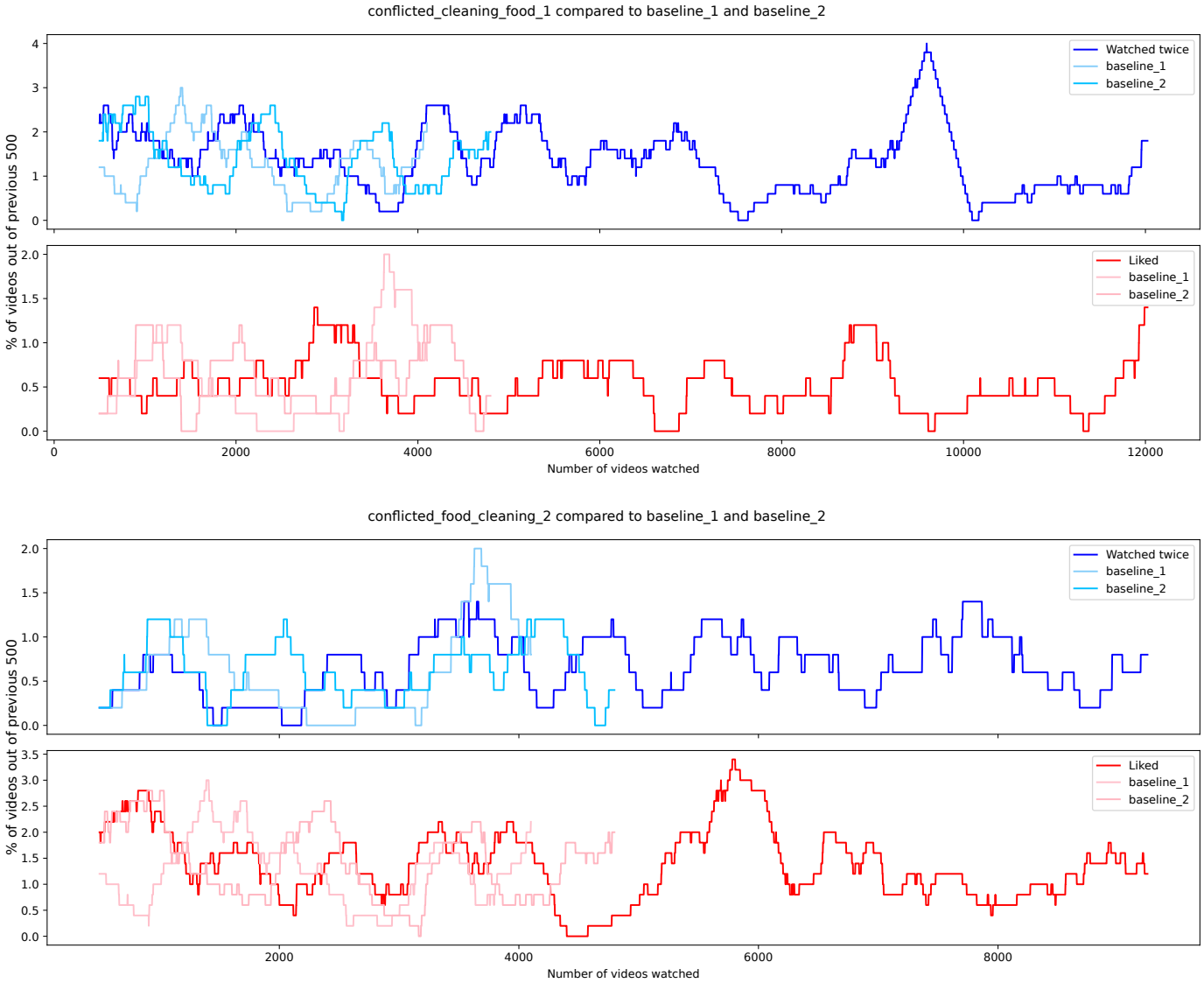

Figure 4.3: Graphs of the rolling sum of videos related to the watched and liked interests of two of the "conflicted" experiments, compared to the rolling sum of videos related to those same interests observed by the two baseline runs. As explained in table 3.1, the first interest in the name of the experiment is the interest which is liked, and the second is the one which is watched twice. By comparing the dark blue and red lines to the lighter lines, we can see here that neither of the "conflicted" runs saw clearly more of the relevant content than the baselines.

5. Conclusions

AI systems are growing in capabilities faster than many have expected, as exemplified, for example, by Google's recent Pathways Language Model (PaLM) [25]. Our capacity to understand and govern them needs to match that growth. Despite the fact that sock puppet audits can be useful (Section 5.1), I have come to believe that they cannot scale to meet that need (Section 5.2) and have many important limitations (Section 5.3). Thus, audits with more internal access are needed (Section 5.4). Finally, I note that many of the problems identified by audits are in act unsolved, and that further research on how to make AI systems aligned with human interests is needed to complement better auditing and governance (Section 5.5).

5.1 Sock puppet audits can be useful

My experiments, described in Chapter 4, showed that personalization did not seem to be as nearly as intense at the time of my experiments as what is commonly believed to be the case, or was observed by the WSJ investigation. Thus, my results support two positive claims about the value of sock puppet audits. Firstly, that sock puppet audits can provide rigorous, novel information about these systems. And secondly, that the WSJ investigation likely led TikTok to make genuine changes their algorithm, and thus that sock puppet audits can affect real change.

It is commonly believed, at least in the U.S. [8] and in Finland [9], that TikTok's recommender system can and does personalize content strongly, accurately, and quickly. The WSJ investigation showed that this perception was largely accurate. But folk theories of algorithms can be substantially inaccurate [26], so rigorous research testing them is valuable. And in this case, that very research made its own conclusions inaccurate, so my newer results provide a useful corrective the now-outdated folk perceptions that it reinforced.

Compared to other types of unauthorized audits, sock puppets also have important advantages in terms of realism and the ability to run controlled experiments. Due to the minimal personalization observed, not was much gained from this in the case of my investigation, but the advantage is real none the less.

My results also lend support the idea that unauthorized external audits can precipitate changes. My experiments were conducted in November and December 2021 (see Table 4.1). TikTok announced changes to their algorithm in December [15]*, likely in response to the WSJ's reporting, as I discussed in Section 2.1. However, as will be discussed in Section 5.3, it is unclear to what extent TikTok's changes really addressed key problems.

5.2 Sock puppet audits cannot scale

As discussed in Section 3.1, sock puppets are relatively challenging to to implement, and their unreliability makes running experiments with them labor intensive. This is due to the fact that websites and apps are actively hostile to bots, and thus significant human assistance and oversight is often required to ensure that the experiments run successfully. These websites and apps also change frequently, often breaking the sock puppet system.

There are numerous massively impactful algorithmic systems in use around the world, and they are constantly changing. To truly understand their impacts, sock puppet experiments would need to be run continuously, in numerous countries, impersonating numerous types of users.

Faced with this, one is tempted to call for a significant technical effort to develop tools to make running these experiments easier. However, such tools could likely be misused for less legitimate uses of social media bots. Similar issues apply to changing the apps or websites themselves to be more bot-friendly.

Other types of unauthorized audits, such as collecting data about content seen by real users as in the Citizen Browser project [18], are also technically demanding, but somewhat more scalable. These can be valuable for getting rough information about what is happening on social media, but cannot tell us about the role of the algorithm in that.

5.3 Unauthorized audits have important limitations

While unauthorized audits can be valuable for identifying problematic behavior, they have significant limitations that make it difficult for them to accurately diagnose prob-

*TikTok's announcement says "we're testing ways to avoid recommending a series of similar content", so it is unclear when exactly they made which changes, but at least some were likely in place in December.

lems, and even more so to identify realistic solutions.

If we take the WSJ's TikTok reporting as an example For example, the WSJ's reporting points out rabbit holes of harmful content, such as content seen as reinforcing mental illness [8] and encouraging eating disorders [14], as a problem on TikTok. A question remains about how the recommender system could be adjusted to avoid this. The WSJ's investigators were not able to say anything in detail about why the system behaved as it did, and how it could be changed to avoid this behavior, without hampering the recommender system's ability to effectively personalize content seen by users in more innocuous cases *.

Pointing out a problem is not enough to show that the most obvious fix would actually result in a system that would be better for users on net. To determine whether the system should be taken offline or replaced requires an understanding of the alternatives and their impacts, which external investigators typically lack. Having access to the company's engineers, and internal data would be necessary for understanding the space of possible solutions, and their likely net impacts on a broad range of users.

In this case, the WSJ's focus on "rabbit holes" led TikTok to increase variability in the feed, and likely also to further bias the recommender system in favor of popular content, in addition to probably suppressing the particular topics identified in the WSJ's reporting as problematic. However, this ad hoc solution does not really bring us closer to a robust solution to the problem of users obsessively focusing on harmful content.

These changes likely made the recommender system "worse" for many people, especially those with interests differing from the mainstream. Not letting users go down rabbit holes will also undermine users' ability to go down delightfully strange rabbit holes leading to unique communities, and is likely associated with mental health costs of its own.

It is also unclear if the problem was really with the recommender system. If this content is genuinely harmful, this appears to be primarily a problem of moderation, not a problem with the recommender system. Making sure that content encouraging eating disorders is balanced with cute cat videos may be an improvement, but it is not really a solution. In the case of depression-related content, the WSJ's sock puppet was completely exclusively interested in that type of content. A real user like this

*One might think that identification of solutions is out of scope for this kind of audit or criticism. However, such work typically does make implicit claims about desirable behavior, and thus makes an implicit claim about how the system should be changed. Thus the authors of such work have a responsibility to consider whether making that change would be net positive. In the case of the WSJ's TikTok investigation, the implicit claim was that there was something concerning about these "rabbit holes", and TikTok responded to that, as discussed below.

would, if their feed was "improved" to include more viral dances, likely move away from the algorithmically curated "for you" page (FYP), opting instead to e.g. only follow accounts posting depressive content, and only browsing their following-page.

5.4 Audits with better access are necessary

Authorized external audits with better access would have a far greater chance of enabling positive change, because they could meaningfully account for the possibilities and constraints faced by the operators of algorithmic systems, and could directly assess the development processes shaping theses systems. Access to internal data would also greatly improve the ability of auditors to assess the net real world impacts of these systems.

When studying such quickly changing systems, it may be more meaningful to assess the development practices and processes giving rise to the system, rather than merely scrutinizing a particular version of a system. Indeed, existing auditing practices in safety critical fields often focus more on assessing processes than individual outputs [27, 28].

5.5 Audits need solutions

Even a combination of unauthorized audits for identifying possible problems, and deeper audits for more precise diagnosis and solution generation, could not overcome the problem that we do not, as a society and as a scientific community, know how to genuinely align AI systems with the interests of individual users, much less society as a whole. As long as these systems blindly optimize for naive metrics like time-on-site, problems will arise, and increasingly so as these systems grow more capable of solving that optimization problem.

Thus there is a need for developing a more coherent theory of the problems associated with algorithmic systems, and a more systematic search for potential solutions. Without an overarching theory of the causes of these problems, it is easy to end up barking up the wrong tree, or even pushing for changes that would be net harmful in the end. In Chapter 1 I attempted to sketch one such theory, focusing on the type of human feedback the system responds to. In practice however, the experiments described in this thesis are not as focused on that theory as they could have been, due to the pressures of journalistic work to simply get out an interesting story.

Building AI systems genuinely aligned with the interests of human beings is a challenging scientific problem that we cannot expect engineers at tech companies to spontaneously solve. Even if deeper audits could provide more insight into the real

world effects of systems, we do not necessarily know how to make systems with better effects. There is much room for meaningful technical research in this direction [29]. Indeed, if we had the ability to make systems more conducive to users' flourishing, tech companies might even switch to them out of self interest. And even if not, reporters and researchers would still stand a much better chance of meaningfully changing things for the better.

As we saw in the case of the WSJ TikTok investigation discussed above, simply identifying that an algorithmic system exhibits certain problematic behaviors given particular inputs, without a clear articulation of the problem, much less a solution, will likely lead operators of algorithms to implement changes that may not effectively address the problem, or, if the problem was incorrectly identified in the critique, may address the wrong problem entirely.

Ultimately, to govern AI well and build a flourishing future for all, we need to combine a wide-ranging search for problematic behavior with thoughtful analyses of problems, and an ambitious effort to develop meaningful solutions.

Bibliography

[1] A. Hannak, "Personalization in online services measurement, analysis, and implications," Ph.D. dissertation, Northeastern University, 2016. [Online]. Available: http://hdl.handle.net/2047/D20235425

[2] M. H. Ribeiro, R. Ottoni, R. West, V. A. F. Almeida, and W. Meira, "Auditing radicalization pathways on YouTube," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 131–141. [Online]. Available: http://doi.org/10.1145/3351095.3372879

[3] M. Faddoul, G. Chaslot, and H. Farid, "A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos," *arXiv:2003.03318 [cs]*, Mar. 2020, arXiv: 2003.03318. [Online]. Available: http://arxiv.org/abs/2003.03318

[4] M. Ledwich and A. Zaitsev, "Algorithmic extremism: Examining YouTube's rabbit hole of radicalization," *First Monday*, vol. 25, no. 3, Feb. 2020. [Online]. Available: https://firstmonday.org/ojs/index.php/fm/article/view/10419

[5] J. Angwin, J. Larson, L. Kirchner, and S. Mattu, "Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk," *ProPublica*, Apr. 2017. [Online]. Available: https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk?token=pc1OkzviktiLco0hJY5BDRPpI44H_bKV

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica*, May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=HM_vaeiiqqmnZ9h19EC5CdgRYKiACk6M

[7] J. Valentino-DeVries, J. Singer-vine, and A. Soltani, "Websites Vary Prices, Deals Based on Users' Information," *Wall Street Jour-*

nal, Dec. 2012. [Online]. Available: https://online.wsj.com/article/SB10001424127887323777204578189391813881534.html

[8] R. Barry, G. Wells, J. West, J. Stern, and J. French, "How TikTok Serves Up Sex and Drug Videos to Minors," *Wall Street Journal*, Sep. 2021. [Online]. Available: https://www.wsj.com/articles/tiktok-algorithm-sex-drugs-minors-11631052944

[9] S. Varpula, "Analyysi | Tiktokin videovirtaan ilmestyi video, joka tiesi minusta pelottavan paljon. Mikä tekee sovelluksen algoritmista niin hyvän?" *Helsingin Sanomat*, Jun. 2021. [Online]. Available: https://www.hs.fi/visio/art-2000008072138.html

[10] Z. Tufekci, "Opinion | YouTube, the Great Radicalizer," *The New York Times*, Mar. 2018. [Online]. Available: https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

[11] J. Vehkoo, "Tiktokin algoritmi ei osaa lukea ajatuksiasi," *Long Play*, Dec. 2021. [Online]. Available: https://www.longplay.fi/sivuaanet/tiktokin-algoritmi-ei-osaa-lukea-ajatuksiasi

[12] S. Clarke and J. Whittlestone, "A Survey of the Potential Long-term Impacts of AI," in *Proceedings of 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, arXiv:2206.11076 [cs]. [Online]. Available: http://arxiv.org/abs/2206.11076

[13] R. Barry, J. West, and G. Wells, "Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires," *Wall Street Journal*, Jul. 2021, section: Video. [Online]. Available: https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/6C0C2040-FF25-4827-8528-2BD6612E3796

[14] T. D. Hobbs, R. Barry, and Y. Koh, "'The Corpse Bride Diet': How TikTok Inundates Teens With Eating-Disorder Videos," *Wall Street Journal*, Dec. 2021. [Online]. Available: https://www.wsj.com/articles/how-tiktok-inundates-teens-with-eating-disorder-videos-11639754848

[15] TikTok, "An update on our work to safeguard and diversify recommendations," Dec. 2021. [Online]. Available: https://newsroom.tiktok.com/en-us/an-update-on-our-work-to-safeguard-and-diversify-recommendations

[16] J. Bandy, "Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits," *Proceedings of the ACM on Human-Computer Interaction*,

vol. 5, no. CSCW1, pp. 74:1–74:34, Apr. 2021. [Online]. Available: http://doi.org/10.1145/3449148

[17] K. Papadamou, S. Zannettou, J. Blackburn, E. de Cristofaro, G. Stringhini, and M. Sirivianos, "'How over is it?' Understanding the Incel Community on YouTube," *Proceedings of the ACM on Human-Computer Interaction*, Oct. 2021, publisher: ACM PUB27 New York, NY, USA. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3479556

[18] "The Citizen Browser Project—Auditing the Algorithms of Disinformation – The Markup," Oct. 2020. [Online]. Available: https://themarkup.org/citizen-browser

[19] C. Lecher and L. Yin, "One Year After the Capitol Riot, Americans Still See Two Very Different Facebooks – The Markup," Jan. 2022, section: Citizen Browser. [Online]. Available: https://themarkup.org/citizen-browser/2022/01/06/one-year-after-the-capitol-riot-americans-still-see-two-very-different-facebooks

[20] C. Faife and A. Ng, "After Repeatedly Promising Not to, Facebook Keeps Recommending Political Groups to Its Users – The Markup," Jun. 2021, section: Citizen Browser. [Online]. Available: https://themarkup.org/citizen-browser/2021/06/24/after-repeatedly-promising-not-to-facebook-keeps-recommending-political-groups-to-its-users

[21] J. Asplund, M. Eslami, H. Sundaram, C. Sandvig, and K. Karahalios, "Auditing Race and Gender Discrimination in Online Housing Markets," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 24–35, May 2020. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/7276

[22] J. Bandy and N. Diakopoulos, "#TulsaFlop: A Case Study of Algorithmically-Influenced Collective Action on TikTok," *CoRR*, vol. abs/2012.07716, 2020, arXiv: 2012.07716. [Online]. Available: https://arxiv.org/abs/2012.07716

[23] D. Klug, Y. Qin, M. Evans, and G. Kaufman, "Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm," in *13th ACM Web Science Conference 2021*, ser. WebSci '21. New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 84–92. [Online]. Available: https://doi.org/10.1145/3447535.3462512

[24] C. Faife, "Facebook Rolls Out News Feed Change That Blocks Watchdogs from Gathering Data – The Markup," Sep. 2021, section: Citizen Browser.

[Online]. Available: https://themarkup.org/citizen-browser/2021/09/21/facebook-rolls-out-news-feed-change-that-blocks-watchdogs-from-gathering-data

[25] S. Narang and A. Chowdhery, "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance," Apr. 2022. [Online]. Available: http://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

[26] M. A. DeVito, D. Gergle, and J. Birnholtz, ""Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, May 2017, pp. 3163–3174. [Online]. Available: https://doi.org/10.1145/3025453.3025659

[27] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensbold, C. O'Keefe, M. Koren, T. Ryffel, J. B. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. Ó hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung, "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," *arXiv:2004.07213 [cs]*, Apr. 2020, arXiv: 2004.07213. [Online]. Available: http://arxiv.org/abs/2004.07213

[28] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 33–44. [Online]. Available: https://doi.org/10.1145/3351095.3372873

[29] J. Stray, I. Vendrov, J. Nixon, S. Adler, and D. Hadfield-Menell, "What are you optimizing for? Aligning Recommender Systems with Human Values," *arXiv:2107.10939 [cs]*, Jul. 2021, arXiv: 2107.10939. [Online]. Available: http://arxiv.org/abs/2107.10939

A. Sock puppet configuration details

High level descriptions of the configurations of the sock puppets were given in Table 4.1, the exact tags that constituted the different interests of the sock puppets, and the methods by which those were compiled, are listed here in Table A.1.

The names and ages of the sock puppet accounts were determined relatively arbitrarily. Ages were somewhere between 18-28, and the names of the accounts were of the form firstname_lastname, with the first and last names randomly picked from a list of the most common first and last names in Finland, with an equal chance of picking male or female first names.

Interest	Keywords & Tags	Created by
School	koulu, opiskelija, lukio, opiskelu, yläaste, alaaste, opettaja, läksyt, kotitehtävät, oppilas, välitunti, #koe	I checked the descriptions of the first 40 videos returned by searching for content tagged #koulu (school) for school related keywords and tags.
Politics	#hallitus, #politiikka, #oikeisto, #vasemmisto, #sdp, #demarit, #perussuomalaiset, #persut, #kokoomus, #keskusta, #kepu, #vihreät, #vasemmistoliitto, #rkp, #kristillisdemokraatit, #kd	I unsystematically looked at what tags were used by politics related content I came across on TikTok.
Cleaning	#siivous, #siivousvinkki, #cleaningtiktok, #cleanfreak, #housekeeping, #cleantok, #cleanhack, #siivousniksi	I searched for videos tagged #siivous (cleaning) and picked out other cleaning related tags those videos used.
Cooking	#baking, #leivonta, #reseptiktok, #keitto, #recipe, #tiktokfood, #ruoka, #ruokaa, #reseptit, #tiktokruoka, #easyrecipe, #reseptitiktok	I searched for with the first of those tags that I came across, and picked out other food and cooking related tags used by those videos.
Depression	#depression, #anxiety, #sad, #pain, #heartbreak	These were the tags mentioned in WSJ's video about their investigation [13].*
Frisbeegolf	#frisbeegolf, #discgolf, #frisbee	I looked at videos tagged #frisbeegolf and picked out frisbeegolf related tags that they used.

Table A.1: The exact keywords and tags that were used to detect videos relating to particular interests. Tags preceded by # were only considered interesting if they appeared verbatim as tags, other strings in the keywords and tags column were considered interesting if they appeared as substrings of the description.