When ‘For You’ Isn’t For You: Measuring User Agency in TikTok’s Algorithmic Feed

Levi Kaplan, Devin Patel, Nicole Gerzon, Alan Mislove, Piotr Sapiezynski
Northeastern University, Boston MA
{kaplan.l, patel.devin1, gerzon.n, p.sapiezynski}@northeastern.edu, amislove@ccs.neu.edu

Abstract

The short-form video-sharing service TikTok has become an important platform in the social media landscape, with much of its popularity owed to its algorithmically-driven “For You Page” (FYP). This feature serves as the “home screen” for the platform and provides a personalized feed of content for each user. Unlike other social media services—where new users start their journey by explicitly signaling whom they choose to friend or follow—the TikTok FYP algorithm instead begins making inferences based on implicit signals, such as how long they watch particular videos. As a result, users have less explicit control over what content they see, and concerns have been raised about the impact on users (e.g., the delivery of potentially harmful content).

In this work, we investigate the extent to which users have control over the content they see on the FYP on TikTok. We first develop novel techniques to study the TikTok mobile app, introducing a new avenue for conducting controlled experiments that enable us to send both explicit and implicit signals on the app. We then use these techniques to study the FYP algorithm based on accounts we control. We find that the FYP algorithm is sensitive to both types of signals, changing the amount of personalized content the account sees. However, we find that users may have difficulty convincing the FYP algorithm to stop showing content the user wishes to no longer see: the most effective explicit signal—marking a video as ‘Not Interested’—is unintuitively buried in the interface. Worse, we find that once accounts cease to indicate disinterest in a topic, many find their feeds dominated by such content again.

1 Introduction

TikTok is the most popular short-form video sharing service in the world (Statista 2024), with over one billion active users each month worldwide (Chew 2024). Much of TikTok’s popularity stems from its “For You Page” (FYP), an algorithmically curated, personalized feed of videos that serves as the homepage for the platform (Hern 2022). While an algorithmic feed of content is not unique to TikTok, their algorithm is said to be highly addictive due to how quickly and accurately it infers user interests (Smith 2021), even to the point of being called “creepy” (Cummins 2022). TikTok has given broad indications of how the FYP algorithm selects content, stating that it often uses *implicit* user signals (e.g., watching a video to the end) in addition to *explicit* user signals (e.g., “liking” a video). Prior work found that following users (an explicit signal) and watch time (an implicit signal) had the greatest influence on the feed (Boeker and Urman 2022). TikTok’s ability to use implicit signals in the algorithm is underscored by their documentation, which states that “the best way to curate your For You feed is to simply use and enjoy the app” (TikTok 2020).

This focus on using implicit signals for recommendation raises concerns about user agency, or users’ ability to control the content they are shown (WSJ 2021). This has come up in discussions of TikTok’s impact on younger users (McCashin and Murphy 2023), creation of filter bubbles (Hunnego 2024; Chen 2023), spread of misinformation (Baumel et al. 2021), and sharing of potentially harmful content (Pryde and Prichard 2022; Liu 2021; Simpson and Semaan 2021; Weimann and Masri 2023). On platforms that rely heavily on explicit signals for personalization, users who are no longer interested in a certain type of content can simply unfollow or unfriend others who publish on that topic. On platforms that rely heavily on implicit signals—such as TikTok—the actions that users should take, and the effectiveness of those actions, are less clear. To provide user agency and control, TikTok has introduced various mechanisms for adjusting their feed, such as refreshing the For You page, using keyword filters to block specific content, and marking videos as ‘Not Interested’ (TikTok 2021). However, the effectiveness of these tools and the extent to which they enable user control is not fully understood.

In this paper, we focus on user agency in influencing TikTok’s FYP algorithm. We aim to understand how quickly the algorithm personalizes content, and how effective the controls the platform provides to users are at “de-personalizing” it. Unfortunately, studying TikTok presents a number of technical challenges. Existing audits have primarily investigated TikTok through its web interface (Boeker and Urman 2022; Vombatkere et al. 2024; Zeng and Abidin 2023). Around two thirds of traffic to TikTok comes from mobile devices (Semrush 2025), and the website has until very recently lacked many features present in the mobile app (TikTok 2025). Additionally, to the best of our knowledge, it is unknown whether the TikTok’s behavior is the same across web and mobile devices, underscoring the importance of investigating the method chosen by most users.

To close this gap, we first develop infrastructure for instrumenting the mobile app and collecting the data it exchanges with the TikTok servers. We further introduce a novel methodology of sock-puppet accounts called "cloning", using which we can test counter-factual scenarios on accounts with identical platform-use histories. This enables us to ask "what if?" questions and better understand how each action influences the FYP algorithm.

We apply these techniques to study user agency in controlling the FYP algorithm, focusing on content with three topics: cooking, fitness, and sports betting. We find that both implicit and explicit negative signals are effective at reducing the amount of topic content delivered to the FYP, and that most of the time (but not always), explicit signals (e.g., marking a video as 'Not Interested') are more effective at reducing unwanted content than implicit signals (e.g., skipping a video). Finally, we find that the FYP algorithm can often "relapse": many accounts which cease to express disinterest and begin watching topical videos again can see their feeds dominated by such videos. This behavior is more common for cooking and fitness than for sports betting.

Overall, our paper makes three contributions:

1. *Novel auditing techniques.* Previous audits on TikTok have primarily focused on the TikTok web interface, which has lacked key features of the mobile app and is used by a minority of the TikTok user base. We present a comprehensive set of techniques for auditing the mobile app through emulator control, mobile app modification, and network traffic interception, providing researchers with the ability to run numerous controlled experiments.
2. *Novel techniques for cloning accounts.* We introduce a novel auditing technique that involves simulating network traffic sent to the TikTok algorithm to precisely duplicate the history of a given account. We demonstrate that the duplicated accounts are treated indistinguishably by the FYP algorithm, enabling "what if?" questions.
3. *Studying user agency in the FYP algorithm.* Using the novel methodologies above, we conduct interventional studies to understand the exact effect that user controls have on the resulting FYP. We show that explicit signaling is often more effective than implicit signaling, and that in many cases, if a user re-engages with content they previously signaled disinterest in, the platform reverts to consistently showing that content.

The remainder of this paper is organized as follows. We provide background and overview related work in Section 2, introduce our methodology in Section 3, and detail our results in Section 4. A concluding discussion is in Section 5.

2 Background and Related Work

We now provide background and describe related work.

2.1 TikTok and Prior Audits

Prior work has highlighted the importance of algorithmic audits on recommendation systems, as they have a serious impact on the information users can access (Bandy 2021). The FYP algorithm relies on user activity (TikTok 2020) and therefore requires historical and real-time data to function.

There are a number of different methodological approaches for conducting algorithmic audits (Sandvig et al. 2014), of which at least two have been applied to TikTok: non-invasive user studies (Zannettou et al. 2024; Vombatkere et al. 2024), and sock-puppet audits (Boeker and Urman 2022; Vombatkere et al. 2024; WSJ 2021). These have focused on both the FYP algorithm and how it categorizes and pushes content (Klug et al. 2021; Scanlon 2021), as well as the behavior of TikTok users and their subsequent outcomes (Boeker and Urman 2022; Vombatkere et al. 2024; WSJ 2021; Zannettou et al. 2024).

This paper provides a new perspective on the former using a sock-puppet audit framework (Sandvig et al. 2014). Unlike previous studies that mainly focused on TikTok's web app (Boeker and Urman 2022; Vombatkere et al. 2024), we develop a novel methodology that can study TikTok's mobile application. Our research further distinguishes itself from prior work by implementing account cloning, which allows for duplicating sock-puppets from the same initial run and offers insight into different app feed control options.

In addition to more formal controlled audits, there is a substantial body of work that explores TikTok algorithmic "folk theories" through qualitative studies of user experience (Karizat et al. 2021; Harris et al. 2023). Folk theories are socially-informed explanations of how an algorithm works in practice, believed to be true by users of the platform (Eslami et al. 2016). Among them, there is scrutiny around whether TikTok's "Not Interested" feature works as intended, with users reporting that TikTok is ineffective in stopping the FYP algorithm from recommending more of the same, potentially damaging, content (DeVito 2022; Harris et al. 2023). We note that even if a formal audit can detect a change caused by the control features, it does not mean that change is commensurate with users' expectations.

2.2 Studies on User Agency

The concept of user agency refers to a user's ability to influence the content they see on a platform. TikTok's FYP is curated by an automated algorithm based on both explicit and implicit signals, such as watching, liking, and sharing recommended content (Boeker and Urman 2022; Kang and Lou 2022; WSJ 2021). Similar recommendation-based platforms, such as YouTube's video recommendations or Facebook's ad delivery algorithm, have historically made it difficult for users to adjust inferred preferences (Geurkink 2019; Panoptykon 2021; Glowacka, Szymielewicz, and Sapiezynski 2023). Even if the features to limit unwanted content exist, issues with effectiveness and usability makes their adoption rare (Gunawan et al. 2021; Gak, Olojo, and Salehi 2022; Habib et al. 2022). This lack of user agency coupled with increased algorithmic personalization has dangerous effects such as amplifying filter bubbles (Woolley and Sharif 2022), spreading misinformation (Tang et al. 2021), and worsening existing harms (Gak, Olojo, and Salehi 2022).

2.3 Fitness Content

With a growing number of people turning to social media platforms for advice on exercise and dieting, "Fitness & Gym" ranks as one of the top categories for content and influencers across social media platforms (Rogers et al. 2022; Stollfuß 2020). While viewing content categorized as "fitness related", viewers are often exposed to content that features excessive dieting, body-shaming, and the glorification of eating disorders (Pryde and Prichard 2022; Liu 2021). Previous research has established that exposure to harmful fitness or dieting content, often demonstrated through popular tags such as "#thinspo" and "#fitspiration", can lead to increased negative body image and dissatisfaction (Norton 2017; Jeronimo and Carraca 2022).

2.4 Sports Betting Content

In 2018, the U.S. Supreme Court ruled that states could legalize sports gambling.[1] Since then, the majority of U.S. states have approved legalization and the sports betting industry now generates over $1B in profit (Matheson 2023). Researchers have studied the harms of sports betting, which has been linked to gambling disorder and addiction (Gainsbury et al. 2015), with negative consequences seen in a higher proportion of younger users (Mestre-Bach et al. 2022). Additionally, sports betting has the potential to be a far more challenging addiction to overcome. While traditional gambling required physical presence in a casino, sports betting is carried out completely virtually, which contributes to higher rates of addiction (Matheson 2023; Gainsbury 2015; Hing et al. 2023), and minimizes the negative feelings of loss resulting from unsuccessful bets (Quick 2024). Prior work has observed that sports betting content is prevalent on TikTok and is actively recommended by the FYP algorithm (Belot 2023).

3 Methodology

We now describe the methodology for running our audit on TikTok to understand user agency in the FYP algorithm.

3.1 Experimental Design

We aim to understand the effectiveness of the tools that TikTok provides in controlling the FYP algorithm. There are a number of mechanisms for users to provide signals to TikTok, including both implicit signals (e.g., a user scrolling past the video, or watching part of the video) or explicit signals (e.g., by clicking on 'Not Interested').[2] We study both types of signals.

Our experiments focus on answering three questions concerning user agency and the FYP algorithm:

1. To what extent does the FYP algorithm personalize content based on user signals?

[1]Murphy v. National Collegiate Athletic Association, No. 16-476, 584 U.S. 453 (2018) [138 S. Ct. 1461]

[2]We note that the 'Not Interested' button is accessible rather nonintuitively via the 'Share' menu or by long-pressing a video, and may not be known to all users.

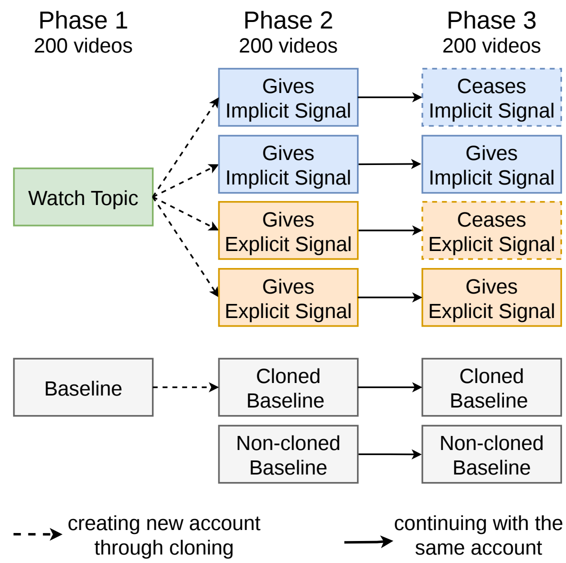

Figure 1: Steps for investigating user agency on TikTok. Each 'Phase' occurs sequentially and involves scraping 200 videos from the FYP. Each element in the flowchart represents a different sock-puppet account ran on a different device. Devices in each column are run simultaneously.

2. To what extent does the FYP algorithm respond if the user signals they no longer wish to see certain content?
3. To what extent does the FYP algorithm start showing content if a user—who previously signaled they did not want to see a particular type of content—stops doing so?

We aim to answer these questions for three different types of content, which we refer to as the three 'topics': cooking, fitness, and sports betting. For each topic, we run five separate experiments, each divided into three phases, as shown in Figure 1. The different experimental roles of the sock-puppet accounts are presented in Table 1.

We have intentionally chosen to run each Phase for the length of 200 videos, as it allows enough samples to provide statistical confidence for observed differences in algorithmic behavior without overwhelming TikTok's servers with requests or providing too much negative feedback so as to not receive topic videos in Phase 3.

Phase 1: Initial Personalization In Phase 1, we aim to understand to what extent the FYP algorithm personalizes content. For a given topic, we first signal interest explicitly by searching for videos related to the topic and watching the first 25 videos from the resulting search results. We then browse 200 videos from the FYP, and continue signaling interest implicitly by fully watching any videos that happen to correspond to the topic of the experiment (and skipping all other videos).[3] We refer to this as the Watch Topic treatment.

Concurrently, we run a separate Baseline treatment, which does not undergo a similar seeding process and simply skips

[3]Whenever we say a video is skipped throughout the paper, we actually watch the video for a random amount of time between 0.2 and 2 seconds, similar to what a human user would do.

Device	Role	Phase	Description
Watch Topic	Cloned	1	The account whose watch history is cloned by the treatment accounts. First undergoes a 'seeding' process, where it watches 25 videos from the search feed, then scrolls through 200 videos from the FYP, watching all videos classified as related to the topic. Only the videos from the FYP are cloned.
Baseline	Cloned	1	The baseline account whose watch history is cloned by the Baseline Clone account. Does not undergo a seeding process, instead only scrolling through 200 videos without watching any videos.
Cloned Baseline	Baseline	2, 3	Clone of Baseline. Represents a baseline that's watched the same number of videos as the treatment devices.
Non-cloned Baseline	Baseline	2, 3	A fresh account that's not been cloned but runs simultaneously to the treatment devices. Represents a baseline that hasn't watched any videos.
Gives Implicit Signal	Treatment	2, 3	Clone of Watch Topic. Skips all videos regardless of whether they are classified as related to the topic.
Gives Explicit Signal	Treatment	2, 3	Clone of Watch Topic. Watches videos that are classified as related to the topic, then marks those videos as 'Not Interested' afterwards.
Ceases Implicit Signal	Treatment	3	Continuation of 'Gives Implicit Signal'. Changes behavior of the Treatment, switching to again watching videos that are classified as related to the topic.
Ceases Explicit Signal	Treatment	3	Continuation of 'Gives Explicit Signal'. Changes behavior of the Treatment, switching to again watching videos that are classified as related to the topic.

Table 1: Description of account types and their experimental roles.

all videos. This allows us to compare the FYP videos shown to the Watch Topic treatment to what an account that did not signal such interest is shown.

Phase 2: Signaling Disinterest In Phase 2, we aim to understand to what extent the FYP algorithm responds if users signal they no longer wish to see content on a given topic. We test the efficacy of explicit signals (i.e., marking videos on the given topic as 'Not Interested', named Gives Explicit Signal) and implicit signals (i.e., skipping videos on the given topic, named Gives Implicit Signal).

To create accounts for Phase 2, we clone[4] the accounts we created in Phase 1 (both the Baseline and the Watch Topic). The cloning process is described in detail in Section 3.4. For the Watch Topic, we actually clone it into four accounts: two each for explicit and implicit signaling. We do so as we will treat some of these accounts differently in Phase 3. For the Baseline, we clone the account into a Cloned Baseline account for consistency in the experimental setup. Finally, we introduce a new account that has not seen any prior videos, which we call Non-cloned Baseline. This account behaves identically to the baseline account from Phase 1; the only difference is that it is not a clone and has not seen any videos beforehand. We run all of these accounts concurrently for 200 videos.

For statistical significance, we repeat the experiment structure in Figure 1 five times for each of our three topics, each with different accounts, for a total of fifteen runs. We then continue these runs in Phase 3.

[4]We only clone the videos seen in the FYP during Phase 1, and we do not clone the 25 videos watched from the search feed.

Phase 3: Interest Relapse In Phase 3, we aim to see to what extent the FYP algorithm responds if a user stops signaling disinterest in the given topic, and instead starts watching such videos (as in Phase 1). We aim here to understand whether TikTok offers opportunities for a user to 'relapse', reverting to a previous behavior or interest which for a period of time they tried to curb.

For each of the two Gives Explicit Signal and Gives Implicit Signal accounts from Phase 2, we have one switch behavior and start watching videos on the topic; we name these Ceases Explicit Signal and Ceases Implicit Signal. For the other account for each, we have it continue giving explicit or implicit signals as in Phase 2. We are interested in whether we detect a 'relapse' effect, which we define as a statistical difference between the amount of topic content delivered to Gives Implicit (Explicit) Signal and Ceases Implicit (Explicit) Signal. We also continue running the Cloned Baseline and Non-cloned Baseline accounts. We run all of these devices for 200 videos on the FYP feed.

3.2 Auditing the TikTok Mobile App

Implementing the experimental design described in the previous section requires overcoming a number of technical hurdles; we describe our approach below.

Emulating the TikTok App To run TikTok on devices that enable automatic control, we employ Android device emulators running through Android Studio. This allows us to simulate many devices at once and control them using our computer. We use the Android Studio CLI[5] to launch the devices, and Android Debug Bridge (`adb`)[6] to connect to our devices and install the TikTok app. We use UIAutomator2[7] to simulate physical interactions with the devices, in particular swiping and clicking.

Creating Accounts Prior to running an experiment, we create all necessary TikTok accounts (including separate accounts for any baseline and cloned treatments). We create six new accounts for each of our fifteen experiments. All account creation is done automatically and simultaneously per experiment. We first reset the Mobile Advertising ID[8] of our Android emulators to prevent any carry-over effects from TikTok storing information based on advertising ID. To further avoid cross-experiment effects, we also reset the data of our TikTok app. We then automatically create TikTok accounts, using randomly generated email addresses. We monitor and manually solve any Captchas that may come up, which are occasionally required to create a TikTok account. After creating an account, TikTok asks the user whether they want to create an account nickname and whether there are any topics they are interested in. We press 'Skip' for both. Finally, a pop-up appears asking whether we approve TikTok accessing our device's contact list for use in finding other users we may know, which we deny.

Intercepting Network Traffic To run our experiments, we need to collect information for each video shown in the FYP feed. While some of the metadata for a video is shown on screen (such as truncated like, comment, and save counts), other metadata fields such as the ID of the video, the length of the video, and the play count of the video are not as easily accessible. However, this additional metadata is available in the HTML of the video's corresponding web URL. We could gather this data by clicking 'Share' on every video and obtaining the URL (Mousavi, Gummadi, and Zannettou 2024), but doing so may be interpreted as a signal of interest by the FYP algorithm. Instead, we intercept the network traffic of the TikTok app, and develop techniques to parse it and find each video's URL.

Intercepting the app's network traffic requires bypassing TikTok's TLS certificate pinning, which protects the network traffic to and from the device from being intercepted and read. We un-pin the certificate using an open-source script (Musa 2023). We then use the Monster-in-the-Middle technique via *Mitmproxy*[9] to intercept the network traffic and collect it.[10] We configure each Android emulator to send all traffic via a separate HTTP proxy, each with a Mitmproxy instance listening to and saving the network traffic. We then parse this saved network data to find the URL of the video and retrieve the metadata from the HTML.

3.3 Deciding Whether to Watch Videos

Now that we have the ability to control the TikTok app and retrieve all the necessary video metadata, we need to determine *how* our accounts interact with TikTok.

Choosing Topics We chose three interest topics for our research into user agency: cooking, fitness, and sports betting. These topics have different levels of prevalence and present different potential harm to TikTok users. Cooking is highly prevalent, and is likely fairly innocuous. While a user could decide they do not want to see cooking videos for mental health or other reasons, the exposure to cooking content is unlikely to be particularly harmful. Fitness is also highly prevalent, and could be considered potentially harmful by some (e.g., pushing unrealistic body standards (Jeronimo and Carraca 2022; Liu 2021; Norton 2017)). Finally, sports betting is not a highly prevent topic, and is considered potentially harmful by some (e.g., by promoting speculative gambling and risky financial decisions (Mestre-Bach et al. 2022; Gainsbury et al. 2015)). In fact, TikTok has specific policies in their Community Guidelines that limits when content related to body image[11] and gambling[12] can be shared.

Classifying Videos Next, we need to determine whether a given TikTok video is on one of the three topics above. Prior work (Boeker and Urman 2022) has used hashtag matching for such classification, but we opt against this for a few reasons. There is necessarily a selection bias for these hashtags; we do not know all the hashtags that are used on the videos we are interested in watching for a given experiment, and new hashtags can come up frequently or use certain euphemisms or codes to get around potential hashtag suppression by the algorithm (Steen, Yurechko, and Klug 2023; Dawson 2024; Cobb 2017). Additionally, many videos do not have hashtags in their description, meaning we would not be able to determine whether to watch those videos.

Instead, we use ChatGPT 3.5 Turbo,[13] a Large Language Model developed by OpenAI, to perform our classification. Through their API,[14] we provide ChatGPT with the video's description, suggested words (a list of related terms that TikTok automatically generates for videos), and hashtags, as well as the nickname (non-unique user-chosen name) and signature (profile biography) of the user who uploaded the video, and ask ChatGPT whether the video is on topic. We provide the exact prompt in the Appendix, Section A.1.

To understand the accuracy of ChatGPT, we compare ChatGPT's classification to human raters. Specifically, for each topic, we selected 150 on-topic videos by searching for each topic using keywords on TikTok and taking the top 150 results. We also selected 150 random videos from the FYP feed. We shuffled the videos, and gave four of the authors of this paper a binary classification labeling task, presenting the same video information that we provide to ChatGPT. For

[5] https://developer.android.com/tools

[6] https://developer.android.com/tools/adb

[7] https://github.com/appium/appium-uiautomator2-driver

[8] https://support.google.com/admanager/answer/6274238?hl=en

[9] https://mitmproxy.org/

[10] We note that we are only intercepting the data that our emulated device sends and receives, and we are not collecting or modifying any other users' personal data.

[11] https://www.tiktok.com/community-guidelines/en/mental-behavioral-health

[12] https://www.tiktok.com/community-guidelines/en/regulated-commercial-activities

[13] https://platform.openai.com/docs/models/gpt-3-5-turbo

[14] https://openai.com/api/

Target Topic	Keywords	Human Fleiss' Kappa	ChatGPT Acc.	Prec.	Recall	F1
Cooking	cooking, recipes, viral recipes, cooking tips, baking	0.922	0.943	0.966	0.791	0.87
Fitness	fitness, health, exercise	0.937	0.940	0.992	0.885	0.935
Sports Betting	sports betting, parlay, fantasy sports, sports gambling	0.963	0.956	0.992	0.915	0.952

Table 2: Keywords used in the ChatGPT prompt, and related validation scores for classification. On the left, we present the Fleiss' Kappa agreements for our human raters (N=4). Given their high agreement, we compare the human labelers' majority vote to ChatGPT's classification and present the accuracy, precision, recall, and F1 score. We overall see very high scores, justifying the use of ChatGPT for classification.

each video, they were asked a yes/no question whether the video was on the given topic.

To measure inter-rater agreement, we use Fleiss' Kappa, which allows us to correct for random chance agreement when measuring inter-rater agreement across multiple raters (Gisev, Bell, and Chen 2013). Table 2 presents the topics, the keywords used in the prompt, and their corresponding inter-rater agreements. We first examine the inter-rater agreement among the humans; we see that the Fleiss' Kappa values are 0.922, 0.937, and 0.963 for cooking, fitness, and sports betting respectively. These values are all high, and are well-above the threshold of 0.81 presented by Landis and Koch for 'almost perfect' agreement (Landis and Koch 1977). To measure the performance of ChatGPT, we take a majority vote from our four human raters, marking disagreement in the case of a tie. We then analyze the accuracy, precision, recall, and F1 score of the performance of ChatGPT compared to the majority vote. These values are presented in Table 2, for each topic. Overall we find very high values—the only value of potential concern is the recall for the cooking topic, at 0.791. A lower recall here indicates that it is not classifying some cooking videos as related to cooking. This would mean that we provide a lower bound for how personalized a feed is, as it likely contains more cooking content than we report. Given the high precision and accuracy, we believe this approach is sufficient for our classification task.

3.4 Cloning Accounts

Finally, we describe our methodology and validation of our approach to TikTok account cloning. Our goal with cloning accounts is to create copies of an account that will share identical watch/skip history, so as to provide the personalization algorithm with the exact same input, resulting in similar personalization outcomes. Pairs of accounts created this way will then enable us to test the efficacy of different treatments. We note that we do not aim to replicate real human behavior, instead seeking to isolate the effects of negative signaling on the resulting algorithmically delivered feed. Additionally, cloning does not imply that they will be delivered identical feeds, due to the inherent randomness of the TikTok algorithm; only that their past histories will be identical and we could expect that they have similar degree and type of personalization in their feeds. To clone an account, we first record the network traffic that it sends, then alter it such that it looks like it came from the new account, and then replay the relevant content back to TikTok's undocumented APIs, without actually watching any videos from the new account. As we show below, this allows us to effectively spoof TikTok into personalizing the FYP as if the new account actually watched the videos we indicated.

Recording and Generating Traffic By analyzing the intercepted network traffic of the Watch Content account, we found that TikTok sends the watch behavior of an account through several API endpoints.[15] While some of these API endpoints receive data about user behavior as either JSON or form data, the `/aweme/v2/feed` endpoint uses a complex protobuf request body.[16] We performed a static analysis of the TikTok app and discovered that it used Square's Wire library[17] to serialize the data, enabling us to generate request data for arbitrary videos. We amalgamate this process of automatically decompiling Square Wire Java protobuf classes in an open-source and general-purpose decompiler.[18]

Similarly, the `/service/2/app_log/` endpoint receives data in the `application/octet-stream;tt-data=b` format. By reverse-engineering the relevant code responsible for generating these request bodies, we determined that the data sent to this endpoint is compressed data using Zstandard[19] with a custom compression dictionary. We pinpointed that this dictionary is loaded from an AES encrypted file at runtime, and extracted the relevant key and IV to decrypt the file ourselves. Re-implementing this compression process in our framework allows for compressed payloads to be generated directly.

Reproducing Security Headers The ability to generate data payloads alone is insufficient to effectively clone user data on TikTok. This is due to TikTok's use of security headers, primarily concerning a proprietary hashing algorithm `ttEncrypt`. This algorithm hashes the entire HTTP request and appends these hashes in various headers. When the server receives a request, it computes the same hashes

[15] We found that it uses the `/aweme/v1/aweme/stats` `/tiktok/v1/realtime/feedback`, `/aweme/v2/feed`, and `/service/2/app_log/` API endpoints.

[16] Protobuf (short for 'protocol buffers') is a data format used to serialize data in such a way that the resulting format is smaller than equivalent payloads in JSON or XML.

[17] https://github.com/square/wire

[18] https://github.com/lumaaaaaa/protoextract

[19] https://github.com/facebook/zstd

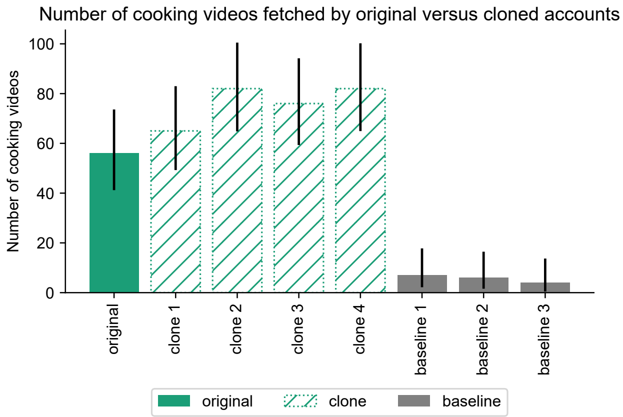

Figure 2: Our Account Cloning technique is successful at transferring personalization from the original account to the cloned account. Bars represent Agresti-Coull 99% confidence intervals.

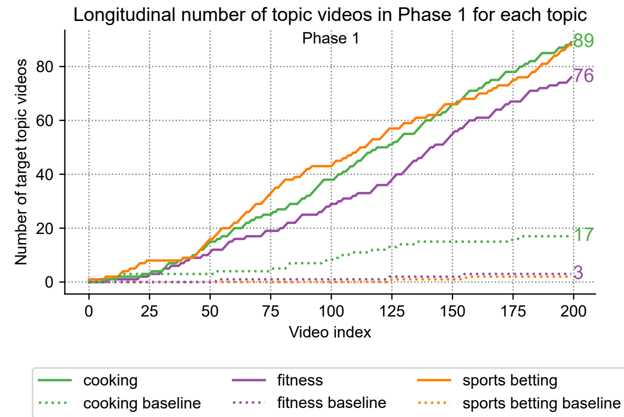

Figure 3: Personalization levels for the three different target topics: cooking, fitness, and sports betting. We see a similar degree of personalization over time between our topics.

for that request and compares them server-side, rejecting the request if the hashes do not match. To properly include these security headers, we use Frida[20] to hook the `ttEncrypt` function at runtime.[21] We can then use the Android device as a 'zombie', hooking the `ttEncrypt` function at runtime to sign the requests we generate in Python, appending valid signature headers to each request sent.

Verifying Cloning Before using the cloning methodology in our experiments, we first verify that the personalization of a cloned account is similar to the original. To do so, we perform a similar process to Phase 1 (Section 3.1) but without first 'seeding' the algorithm, instead watching videos from the default FYP that are related to cooking. This is because our cloning technique is only able to duplicate watching videos from the FYP and cannot replicate searching for videos and watching videos in the search feed. We refer to this account as the original account. Using separate emulators, we then clone the watch history of this account four times; we refer to these as the cloned accounts. If cloning works, then we would expect the subsequent FYP feed delivered to the original account to be similar to those delivered to the cloned account. For reference, we also include three separate baseline accounts with no history and which are not cloned, to measure the fraction of cooking videos that we can expect would appear in the FYP feed randomly.

If we were to take these eight accounts and scrape the FYP feed, we would necessarily be biasing the algorithm. This is because any amount of watch time acts as a signal to the algorithm: if we simply skipped through a set of videos, we would see fewer topic videos over time, and if we instead watched the videos, we would see more topic videos over time. Therefore, to understand whether our cloning methodology works without affecting the algorithm, we make use of another API hidden in the TikTok app. Whenever the FYP feed is fetched by an account, the `/api/v2/feed` endpoint is called. We discovered we can call this endpoint successively and it returns the next N videos from the feed *without* biasing the algorithm via sending whether the video was watched or not. We call this process *fetching* the feed. By fetching the next 200 videos, we can get an idea of the level of personalization in the algorithm at a given moment, without biasing the algorithm by sending negative or positive watch signals from scrolling through the feed.

For all eight accounts, we fetch the next 200 videos from the FYP endpoint using this process. We hypothesize that the distribution of delivery of cooking videos to the original and cloned accounts' feeds will be similar, and that they will be different from the baseline. The results of this experiment can be found in Figure 2, with bars representing the Agresti-Coull 99% confidence intervals. We see that all four clones see a considerably higher quantity of cooking videos than the baselines, and that the clones are not statistically distinguishable from themselves or from the original account, indicating that the personalization from the original device successfully transferred over.

4 Results

We now present the results from our study.

4.1 Phase 1: Initial Personalization

Recall that our goal in Phase 1 is to understand to what extent the FYP algorithm personalizes content. For each of the three topics, we first 'seed' an account by searching for the topic and watching the first 25 videos, and then collect 200 videos on the FYP feed (watching any on-topic videos and skipping all others). We also run a separate baseline account at the same time which skips all videos.

Figure 3 shows the number of topic videos delivered over the course of the experiment (after the 'seeding') for each topic. The x-axis represents the index of the video in the

[20] https://frida.re/

[21] Frida is a general purpose dynamic instrumentation framework that enables hooking into the function in real time, allowing us to export the call to the hashing algorithm directly to Python.

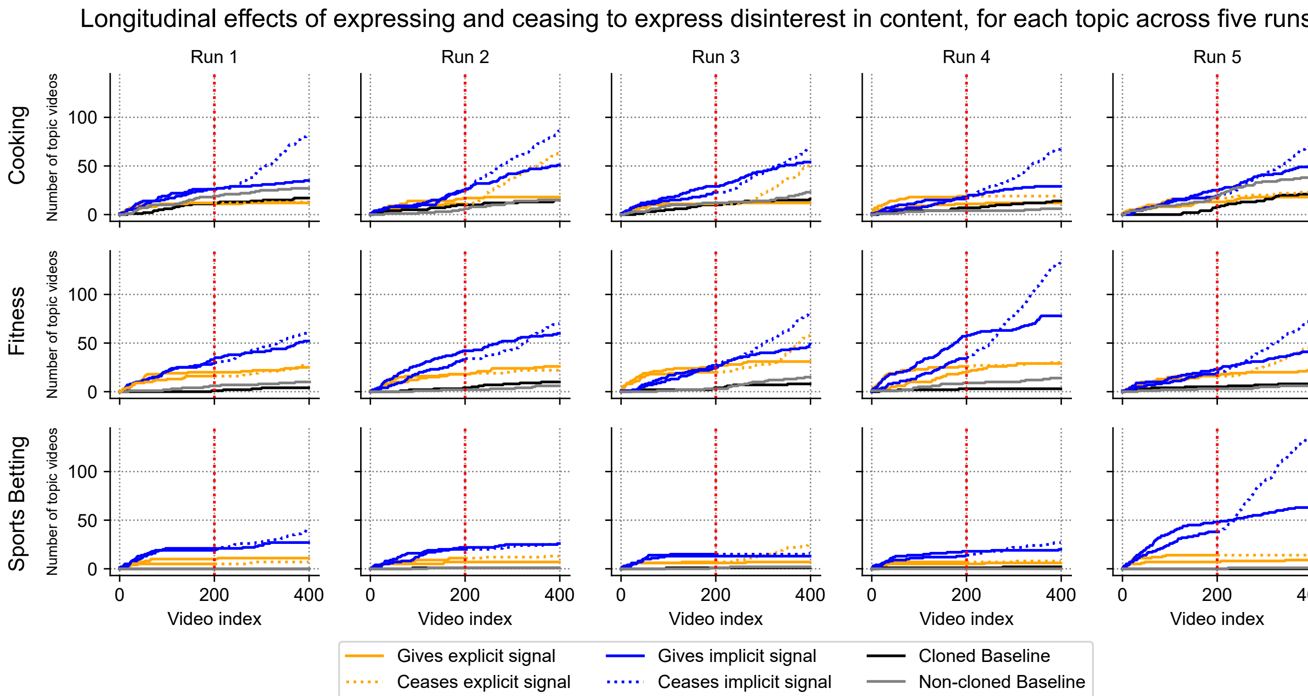

Figure 4: The number of topic videos delivered to our accounts across our fifteen experiments. The results of Phase 2 can be found on the left side of the red dotted line, and the results of Phase 3 can be found to the right side of the red dotted line, for each experiment. Dotted blue (orange) lines represent accounts which cease implicit (explicit) signaling, going back to watching topic content.

FYP feed (from 1 to 200), and the y-axis indicates the number of videos on the topic in the FYP feed at that point. At x=200, the y value therefore represents the total number of target topic videos seen by the device over the phase.

We see that for all topics, the algorithm quickly personalizes the content delivered via the FYP algorithm. We achieved a similar degree of personalization for all three topics: the cooking and sports betting topics saw 89 (44.5%) videos related to their respective topics. The fitness topic saw sightly fewer at 76 (38%). Looking at the topics' prevalence in the baseline, we see 17 videos related to cooking (8.5%), and only three related to fitness and sports betting (1.5%).

4.2 Phase 2: Signaling Disinterest

In Phase 2, we aim to understand the extent the FYP algorithm responds when a user signals—either implicitly or explicitly—that they no longer wish to see the content that they previously engaged with. To that aim, we first run a simple experiment to see whether negative signals impact the FYP algorithm's recommendations at all. We focus on cooking, using six devices that have been cloned from a Phase 1 cooking account: two devices that watch all cooking videos, two that signal disinterest implicitly, and two that do so explicitly. We also include two baseline accounts. To measure statistical significance across the two accounts of each treatment, we add up the total number of on-topic videos seen as the sample proportion and use 400 (200 videos for each account) as the sample size. We then run a two-proportion Z-test at the 99% confidence level. Figure 5 shows that there were fewer target videos in the feeds of accounts who changed their behaviors, compared to those that did not.

The efficacy of the interventions differed, however. On-topic videos constituted 16% of the implicit signaling account's feed, compared 30.5% of the feed of the account which continued watching them (a 47.5% decrease). Explicit

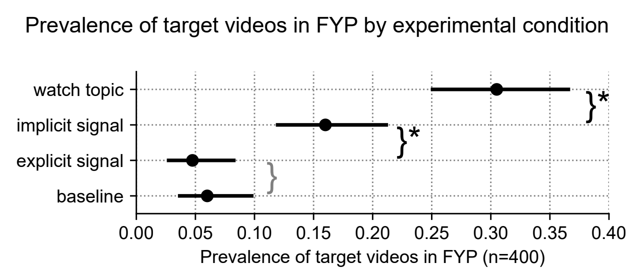

Figure 5: The account which sends an implicit positive signal (watch topic) receives more such videos than the account which sends an implicit negative signal. Sending an explicit negative signal results in even fewer on-topic videos, bringing the FYP close to the non-personalized baseline. The $\star$ symbol in the figure marks which differences in prevalence are significant at the 99% confidence level.

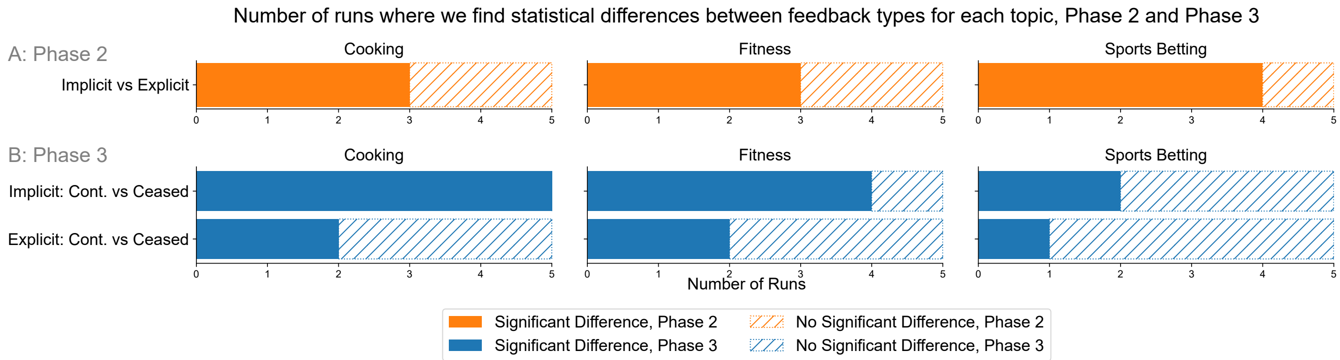

Figure 6: Counts where we find statistical differences for the fifteen experiments across three topics, for Phases 2 and 3. Phase 2 (row A): we typically see statistically more topic videos when implicitly signaling than when explicitly signaling across our three topics. Phase 3 (row B): The cooking and fitness topics consistently relapsed in the implicit case, and relapsed in the explicit case twice. The sports betting topic was found to relapse less often in both the implicit and explicit cases.

signaling led to an even more pronounced decrease at 4.75% (a total decrease of 84.4%). In fact, in this experiment, the difference between the prevalence of the on-topic videos in the feed of the explicit signaling account is not significantly different from that of a non-personalized baseline account.

Having shown that both explicit and implicit signaling leads to measurable effects on one topic (cooking), we now turn to including all three target topics and running Phases 2 and 3 of the experiment presented in Figure 1 five times per topic, for a total of fifteen experimental runs. The number of topic videos delivered to each of our accounts during Phase 2, across all fifteen runs, is shown to the left side of the dotted red line for each of the sub-figures in Figure 4. Overall we see a reduction in the amount of personalized content over time throughout each of our runs for both implicit and explicit signaling, suggesting that both strategies reduce the amount of topic content on the FYP over time.

Whether explicit signaling is more effective than implicit signaling differs across topics and runs, however. Figure 6A tallies the number of runs where a statistical difference was found between the accounts using implicit versus explicit signaling. For each experiment, since we have two accounts for each signaling type, we perform a similar process as before, adding together the total number of topic videos seen for each identical signaling account as the sample proportion, adding up the total number of videos seen (200 per account, for a total of 400) as the sample size, and then running a two-proportion Z-test at the 99% confidence level. For most runs, we see statistically fewer topic videos for the explicit signaling case compared to implicit signaling. However, this is not the case for all runs. Sometimes, we see implicit signaling to be as effective as explicit signaling, considerably reducing the amount of topic content seen by the account; this happens for two out of five runs for cooking and fitness, and one out of five runs for sports betting.

4.3 Phase 3: Treatment Relapse

Finally, recall that in Phase 3 our goal is to understand to what extent the FYP algorithm starts showing content that a user previously signaled disinterest in, after the signaling ceases. Also recall that Phase 3 introduces a change in behavior for one of the two treatment accounts for each treatment condition. For each target topic and type of signaling, one account continues the signaling, while the other switches to watching on-topic videos. We will refer to the accounts that switch behavior as ceases implicit (explicit) signaling and the accounts that continue signaling as continues implicit (explicit) signaling. The number of topic videos delivered to each of our accounts during Phase 3 can be found to the right side of the dotted red line in Figure 4.

We define a 'relapse' as an instance where the ceases implicit (explicit) signaling account sees statistically more topic videos than the corresponding continues implicit (explicit) signaling. We calculate statistical difference using a two-proportion Z-test at the 99% confidence level. Recall that we run these accounts simultaneously—we can compare the behavior of the account that continues its negative signaling to the corresponding account that ceases that signaling and begins watching all topic videos to understand whether the algorithm begins pushing more topic content to that account relative to the account that continues negative signaling.

We run this experiment five times for each topic, continuing the fifteen experiments described in the section above, and measure the number of successful relapses for the implicit signaling and explicit signaling cases. These results can be found in Figure 6B. We detail the results for ceasing implicit signaling and ceasing explicit signaling below.

Ceasing Implicit Signaling We will first explore the results for relapsing in the implicit signaling case. For all three topics, we were able to achieve a *relapse* effect multiple times. For the cooking topic, we triggered a relapse for all five runs. This indicates that implicit signaling was not enough to prevent a relapse for an account that returns to watching that content after negatively signaling implicitly for 200 videos.

For fitness and sports betting, we had runs that did not see an implicit relapse. For fitness this only occurred once, which indicates that implicit negative feedback is likely not sufficient in the majority of cases to avoid seeing that content in the FYP in the future. However, we only saw a relapse in two out of five experiments for sports betting. This indicates that relapsing for that topic is less common. Still, with how sensitive these two topics can be, it is interesting that implicit feedback was not enough in multiple cases to prevent the algorithm from pushing that content again.

Ceasing Explicit Signaling We now turn to the accounts that cease explicit signaling (i.e., marking videos 'Not Interested'), to investigate how common relapsing is for a stronger negative signal. We observe fewer relapses compared to the implicit case; two accounts relapsed for the cooking and fitness topics, and one relapsed for the sports betting topic. This indicates that in our controlled environment explicit signaling is more effective than implicit signaling to reduce future exposure should a user's behavior change, across all topics. However, the fact that we were able to trigger a relapse for all three topics indicates that, for some users, their explicit feedback could be overridden if there is a change in behavior in the future.

5 Discussion

In this work, we examined user agency in the context of the FYP algorithm on TikTok, aiming to understand how explicit and implicit signals can be used to control the content users see. We introduced new methodologies to study user agency in the context of three topics: cooking, fitness, and sports betting. We found that, in the experimental setup we created, both explicit signals (i.e., marking a video as 'Not Interested') and implicit signals (i.e., skipping past videos) are effective in reducing the delivery of that content. Overall, we find explicit signaling to often be more effective when compared to implicit signaling. However, as we explain below, this finding does not necessarily invalidate the reports of real users who experienced inefficacy of these signals.

Unfortunately, when investigating a 'relapse' effect of switching behavior back to watching that content, we find that accounts that used implicit signals were more often able to 'relapse' once such signals ceased. Even when investigating stronger explicit signals, we were still able to trigger a relapse effect for two accounts in the cooking and fitness topics, and one for the sports betting topic.

5.1 Limitations

Relying on machine-controlled sock-puppet accounts allows for precise isolating of different factors that might contribute to the overall user experience on the platform. At the same time, this strength is also the core weakness of the approach. The effects we observe in this work are strictly driven by the limited set of behaviors we simulated: watching videos, skipping videos, and marking videos as 'Not Interested'. Real users might perform many actions we chose not to simulate: liking, sharing, and reporting videos, commenting on them, as well as following and un-following creators. Real users' location might further influence their experience, as can their browsing histories, or being targeted by ads. When real users make a decision on how to react to each video, the signals they consider are different than those available to our sock puppet accounts. Additionally, real users might not be focused on one topic exclusively, or as persistent as our sock-puppet accounts in expressing disinterest.

We chose to focus on a limited number of topics, but there are a wide array of potential interests and real users likely have a wide variety of personalized videos delivered to their feed. In order to isolate the effect of negative signals our sock-puppet accounts did not substitute the topic of disinterest for a new topic of interest. Real users might, in the process of expressing disinterest in a particular topic, be delivered content that they like and start watching it, thereby shaping their feed more positively.

Finally, we are studying the state of the algorithm at a single point in time. TikTok may adjust how effective negative feedback controls are in controling a user's FYP. We nevertheless believe that illustrating how effective these controls are in a controlled setting can help users understand how their actions can impact their experience on the platform.

Because of these limitations we do not claim that real users will experience the same extent of personalization in their feed, equivalent levels of efficacy of user controls, or the rate of "re-personalization" after they cease to use those controls. Our results do not undermine the lived experience of users who did not observe the effects of sending explicit signals of disinterest. In fact, both can be true—the change introduced by expressing disinterest might be statistically significant, while still not matching real users' expectations. We hope that, despite its limitations, our research will inform future work with real users.

5.2 Ethical Considerations

Over the course of this research, we made efforts to ensure ethical data collection and processing. We only collected data from public profiles delivered to the FYP or search feeds, and we did not collect any private user data. We minimized our load on TikTok's servers by only creating the accounts necessary to conduct our experiments. We minimized harm to other TikTok users by only interacting with the platform through watching videos; we did not like videos, comment, or follow creators, to minimize the effect our accounts had on the platform's algorithms. We also avoided watching content promoting illegal or otherwise unethical behavior, choosing topics that represented potentially sensitive but nonetheless allowed content on the platform.

Acknowledgments

This work has been funded in part by the National Science Foundation grant CNS-2318290. We thank the anonymous ICWSM Reviewers for their helpful feedback.

References

Bandy, J. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1): 1–34.

Baumel, N. M.; Spatharakis, J. K.; Karitsiotis, S. T.; and Sellas, E. I. 2021. Dissemination of mask effectiveness misinformation using TikTok as a medium. *Journal of Adolescent Health*, 68(5): 1021–1022.

Belot, H. 2023. TikTok removes video critical of gambling advertising while increasing wagering content. https://www.theguardian.com/australia-news/2023/sep/21/tiktok-removes-video-critical-of-gambling-advertising-while-increasing-wagering-content. Accessed: 2025-01-10.

Boeker, M.; and Urman, A. 2022. An empirical investigation of personalization factors on tiktok. In *Proceedings of the ACM Web Conference 2022*, 2298–2309.

Chen, S. 2023. How Social Media Can Solve the Problem of "Filter Bubbles" Under the NewMedia Algorithm Recommendation Mechanism the Example of Tik Tok. In *2023 2nd International Conference on Social Sciences and Humanities and Arts (SSHA 2023)*, 1284–1288. Atlantis Press.

Chew, S. 2024. TikTok CEO Shou Chew's Opening Statement. https://newsroom.tiktok.com/en-us/opening-statement-senate-judiciary-committee-hearing. Accessed: 2025-01-12.

Cobb, G. 2017. "This is not pro-ana": Denial and disguise in pro-anorexia online spaces. *Fat Studies*, 6(2): 189–205.

Cummins, E. 2022. The Creepy TikTok Algorithm Doesn't Know You. https://www.wired.com/story/tiktok-algorithm-mental-health-psychology/. Accessed: 2025-01-13.

Dawson, S. 2024. *You can't say that on TikTok: cxnsxrshxp, algorithmic (in) visibility, and the threat of representation*. Ph.D. thesis, University of British Columbia.

DeVito, M. A. 2022. How transfeminine TikTok creators navigate the algorithmic trap of visibility via folk theorization. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–31.

Eslami, M.; Karahalios, K.; Sandvig, C.; Vaccaro, K.; Rickman, A.; Hamilton, K.; and Kirlik, A. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*, 2371–2382.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gainsbury, S. M. 2015. Online gambling addiction: the relationship between internet gambling and disordered gambling. *Current addiction reports*, 2(2): 185–193.

Gainsbury, S. M.; Russell, A.; Hing, N.; Wood, R.; Lubman, D.; and Blaszczynski, A. 2015. How the Internet is changing gambling: Findings from an Australian prevalence survey. *Journal of Gambling Studies*, 31: 1–15.

Gak, L.; Olojo, S.; and Salehi, N. 2022. The distressing ads that persist: Uncovering the harms of targeted weight-loss ads among users with histories of disordered eating. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–23.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Geurkink, B. 2019. Our Recommendation to YouTube. https://foundation.mozilla.org/en/blog/our-recommendation-youtube/. Accessed: 2025-01-10.

Gisev, N.; Bell, J. S.; and Chen, T. F. 2013. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3): 330–338.

Glowacka, D.; Szymielewicz, K.; and Sapiezynski, P. 2023. Algorithms of Trauma 2: Stuck in a "doomscrolling trap" on Facebook? The platform will not let you escape. https://panoptykon.org/sites/default/files/2023-12/panoptykon_algorithms-of-trauma-2_case-study-report_dec-2023.pdf. Accessed: 2025-01-10.

Gunawan, J.; Pradeep, A.; Choffnes, D.; Hartzog, W.; and Wilson, C. 2021. A comparative study of dark patterns across web and mobile modalities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–29.

Habib, H.; Pearman, S.; Young, E.; Saxena, I.; Zhang, R.; and Cranor, L. F. 2022. Identifying user needs for advertising controls on Facebook. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–42.

Harris, C.; Johnson, A. G.; Palmer, S.; Yang, D.; and Bruckman, A. 2023. " Honestly, I Think TikTok has a Vendetta Against Black Creators": Understanding Black Content Creator Experiences on TikTok. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–31.

Hern, A. 2022. How TikTok's algorithm made it a success: 'It pushes the boundaries'. https://www.theguardian.com/technology/2022/oct/23/tiktok-rise-algorithm-popularity. Accessed: 2025-01-12.

Hing, N.; Browne, M.; Rockloff, M.; Russell, A. M.; Tulloch, C.; Lole, L.; Thorne, H.; and Newall, P. 2023. Situational features of smartphone betting are linked to sports betting harm: An ecological momentary assessment study. *Journal of Behavioral Addictions*, 12(4): 1006–1018.

Hunnego, M. 2024. *Exploring the Consistency and Variability of Algorithmic Filter Bubbles: A Comparative Analysis of Instagram Reels and TikTok*. B.S. thesis, University of Twente.

Jeronimo, F.; and Carraca, E. V. 2022. Effects of fitspiration content on body image: a systematic review. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, 27(8): 3017–3035.

Kang, H.; and Lou, C. 2022. AI agency vs. human agency: understanding human–AI interactions on TikTok and their implications for user engagement. *Journal of Computer-Mediated Communication*, 27(5): zmac014.

Karizat, N.; Delmonaco, D.; Eslami, M.; and Andalibi, N. 2021. Algorithmic folk theories and identity: How TikTok users co-produce Knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2): 1–44.

Klug, D.; Qin, Y.; Evans, M.; and Kaufman, G. 2021. Trick and please. A mixed-method study on user assumptions about the TikTok algorithm. In *Proceedings of the 13th ACM Web Science Conference 2021*, 84–92.

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Liu, J. 2021. The influence of the body image presented through TikTok trend-videos and its possible reasons. In *2nd International Conference on Language, Art and Cultural Exchange (ICLACE 2021)*, 359–363. Atlantis Press.

Matheson, V. 2023. Sports Gambling. https://www.milkenreview.org/articles/sports-gambling. Accessed: 2025-01-10.

McCashin, D.; and Murphy, C. M. 2023. Using TikTok for public and youth mental health–A systematic review and content analysis. *Clinical Child Psychology and Psychiatry*, 28(1): 279–306.

Mestre-Bach, G.; Granero, R.; Mora-Maltas, B.; Valenciano-Mendoza, E.; Munguía, L.; Potenza, M. N.; Derevensky, J. L.; Richard, J.; Fernández-Aranda, F.; Menchón, J. M.; et al. 2022. Sports-betting-related gambling disorder: Clinical features and correlates of cognitive behavioral therapy outcomes. *Addictive Behaviors*, 133: 107371.

Mousavi, S.; Gummadi, K. P.; and Zannettou, S. 2024. Auditing Algorithmic Explanations of Social Media Feeds: A Case Study of TikTok Video Explanations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1110–1122.

Musa, E. 2023. TikTok SSL Pinning Bypass. https://github.com/Eltion/Tiktok-SSL-Pinning-Bypass.

Norton, M. 2017. Fitspiration: Social Media's Fitness Culture and its Effect on Body Image.

Panoptykon, F. 2021. Algorithms of trauma: new case study shows that Facebook doesn't give users real control over disturbing surveillance ads. https://en.panoptykon.org/algorithms-trauma-new-case-study-shows-facebook-doesnt-give-users-real-control-over-disturbing. Accessed: 2025-01-10.

Pryde, S.; and Prichard, I. 2022. TikTok on the clock but the# fitspo don't stop: The impact of TikTok fitspiration videos on women's body image concerns. *Body image*, 43: 244–252.

Quick, J. 2024. 'I literally can't stop.' The descent of a modern sports fan. https://www.nytimes.com/athletic/5777632/2024/10/14/sports-betting-addiction-problem-fans/. Accessed: 2025-01-10.

Rogers, A.; Wilkinson, S.; Downie, O.; and Truby, H. 2022. Communication of nutrition information by influencers on social media: A scoping review. *Health Promotion Journal of Australia*, 33(3): 657–676.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014): 4349–4357.

Scanlon, K. 2021. The App That Knows You Better than You Know Yourself: An Analysis of the TikTok Algorithm.

Semrush. 2025. tiktok.com Web Traffic Statistics. https://www.semrush.com/website/tiktok.com/overview/.

Simpson, E.; and Semaan, B. 2021. For you, or for "you"? Everyday LGBTQ+ encounters with TikTok. *Proceedings of the ACM on human-computer interaction*, 4(CSCW3): 1–34.

Smith, B. 2021. How TikTok Reads Your Mind. https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html. Accessed: 2025-01-13.

Statista. 2024. Most popular short video platform and features worldwide 1st quarter 2024, average video views. https://www.statista.com/statistics/1466343/short-video-platform/. Accessed: 2025-01-12.

Steen, E.; Yurechko, K.; and Klug, D. 2023. You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media+ Society*, 9(3): 20563051231194586.

Stollfuß, S. 2020. Communitainment on Instagram: Fitness content and community-driven communication as social media entertainment. *Sage Open*, 10(2): 2158244020919535.

Tang, L.; Fujimoto, K.; Amith, M.; Cunningham, R.; Costantini, R. A.; York, F.; Xiong, G.; Boom, J. A.; and Tao, C. 2021. "Down the rabbit hole" of vaccine misinformation on YouTube: Network exposure study. *Journal of Medical Internet Research*, 23(1): e23262.

TikTok. 2020. How TikTok recommends videos #ForYou. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you. Accessed: 2025-01-12.

TikTok. 2021. How TikTok recommends content. https://support.tiktok.com/en/using-tiktok/exploring-videos/how-tiktok-recommends-content. Accessed: 2025-01-14.

TikTok. 2025. New Features Bring TikTok's Handheld Magic to Desktop. https://newsroom.tiktok.com/en-us/new-features-bring-tiktok-magic-to-desktop. Accessed: 2025-05-14.

Vombatkere, K.; Mousavi, S.; Zannettou, S.; Roesner, F.; and Gummadi, K. P. 2024. TikTok and the Art of Personalization: Investigating Exploration and Exploitation on Social Media Feeds. In *Proceedings of the ACM on Web Conference 2024*, 3789–3797.

Weimann, G.; and Masri, N. 2023. Research note: Spreading hate on TikTok. *Studies in conflict & terrorism*, 46(5): 752–765.

Woolley, K.; and Sharif, M. A. 2022. Down a rabbit hole: how prior media consumption shapes subsequent media consumption. *Journal of Marketing Research*, 59(3): 453–471.

WSJ. 2021. Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires. 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.

Zannettou, S.; Nemes-Nemeth, O.; Ayalon, O.; Goetzen, A.; Gummadi, K. P.; Redmiles, E. M.; and Roesner, F. 2024. Analyzing User Engagement with TikTok's Short Format Video Recommendations using Data Donations.

Zeng, J.; and Abidin, C. 2023. '# OkBoomer, time to meet the Zoomers': Studying the memefication of intergenerational politics on TikTok. In *The playful politics of memes*, 93–115. Routledge.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this research question seeks to better understand the recommendation algorithm and user agency for users on TikTok
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. See Sections 3 and 4
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see section 3.1 for a description of how our methodology answers our research questions
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? N/A
   (e) Did you describe the limitations of your work? Yes, the limitations to our methodology are discussed in Section 5.1.
   (f) Did you discuss any potential negative societal impacts of your work? We discuss the potential harms of our research methodology in Section 5.2, and the approach we took to mitigating and minimizing those harms.
   (g) Did you discuss any potential misuse of your work? We do not anticipate any misuse, but we do provide discussion on the inaccuracies of applying these findings outside of their scope in Section 5.1
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Sections 3.2 and 3.4 for information on reproducibility and Section 5.2 for information on scraping public data.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? While the main experimental result was not ML focused, we did use a ChatGPT as part of our methodology to classify videos. Information on our methodology and reproducibility can be found in Sections 3 and 4. We validated our methodology in Section Section 3.3, and provide inter-rater agreements using Fleiss' Kappa for the four human annotators and the accuracy, precision, recall, and F1 scores for ChatGPT measured against human classification majority vote in Table 2. Section A.1 in the Appendix includes the prompts we gave the LLM. Section 3.3 discusses the reasoning behind using an LLM and why we feel confident in that decision.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, we use multiple different tools throughout the experiment design that is cited in Sections 3.2, 3.3, and 3.4.
   (b) Did you mention the license of the assets? The assets we use are open-source tools or libraries that do not require licenses.
   (c) Did you include any new assets in the supplemental material or as a URL? No.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Section 5.2.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? We only collect data that has been

posted publicly, and we do not collect private or personally identifiable information about any users or creators on the platform. We avoid watching content that promotes illegal/unethical behavior, and only engage with specific categories of content that are allowed on TikTok. The sensitive topics we look at in this paper may be considered offensive, as discussed in Sections 2.3 and 2.4, however they do not violate any community or content guidelines. This is further discussed in Section 5.2.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? No new datasets are being curated or released,

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? N/A

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? N/A

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

(d) Did you discuss how data is stored, shared, and deidentified? N/A

A Appendices

A.1 ChatGPT Prompts

We used the following prompt structure, with the target topic and keywords presented in Table 2 substituted in:

"You are a classifier tasked with determining whether the given content has anything to do with [keywords]. Given a list of: the user who posted a video and a brief description of them; the video's description; a list of related words; and a list of hashtags, you classify whether the video is related to [topic]. You only respond with 'Yes' if you think it is, or 'No' if not.

Description: [...], Hashtags: [...], Suggested Words: [...], Nickname: [...], Signature: [...]"

A.2 Dates for Experiments

Table A1 presents the dates when each Phase of each of our experiments were run. We ran Phase 1 once for each topic. We ran Phases 2 and 3 five times for each topic, for a total of 15 runs.

Topic	Phase 1	Phases 2/3									
		Run 1		Run 2		Run 3		Run 4		Run 5	
		Phase 2	Phase 3	Phase 2	Phase 3	Phase 2	Phase 3	Phase 2	Phase 3	Phase 2	Phase 3
Cooking	3/17/25	3/27/25	3/27/25	3/31/25	3/31/25	4/11/25	4/11/25	4/15/25	4/15/25	4/17/25	4/17/25
Fitness	3/17/25	4/4/25	4/4/25	4/9/25	4/9/25	4/16/25	4/16/25	4/21/25	4/21/25	4/22/25	4/22/25
Betting	3/17/25	4/10/25	4/10/25	4/10/25	4/10/25	4/14/25	4/14/25	4/16/25	4/16/25	4/18/25	4/18/25

Table A1: The dates when our various experiments were run.