Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Auditing Algorithmic Explanations of Social Media Feeds: A Case Study of TikTok Video Explanations

Sepehr Mousavi[1], **Krishna P. Gummadi**[1], **Savvas Zannettou**[2]
[1]Max Planck Institute for Software Systems, Saarbrücken, Germany
[2]TU Delft, Delft, The Netherlands
smousavi@mpi-sws.org, gummadi@mpi-sws.org, s.zannettou@tudelft.nl

Abstract

In recent years, user feeds on social media platforms have shifted from simple, chronologically ordered content posted by their network connections (i.e., friends) to opaque, algorithmically ordered and curated content. This shift has led to regulations that require platforms to offer end users greater transparency and control over their algorithmic recommendation-based feeds. In response, social media platforms such as TikTok have recently started explaining why specific videos are recommended to end users. However, we still lack a good understanding of how these explanations are generated and whether they offer the desired transparency to end users. In this work, we audit explanations provided on short-format videos on TikTok. We collect a large dataset of short-format videos and explanations provided by TikTok (when available) using automated sockpuppet accounts. Then, we systematically characterize the explanations, focusing on their accuracy and comprehensiveness. For our assessments, we compare the provided explanations with video metadata and the behavior of our sockpuppet accounts. Our analysis shows that some generic (non-personalized) reasons are always included in explanations (e.g., "This video is popular in your country"), while at the same time, we find that a large number of provided explanations are incompatible with the behavior of our sockpuppet accounts; e.g., an account that made zero comments on the platform, was presented with the explanation "You commented on similar videos" in 34% of all recommended videos. Overall, our audit of TikTok video explanations highlights the need for more accurate, fine-grained, and useful explanations for the end users. We will make our code and dataset available to assist the research community.

1 Introduction

As you browse through your social media feeds, be it TikTok videos, Instagram Reels, YouTube Shorts, Facebook newsfeed, or Twitter timeline, have you ever wondered how or why the next content instance is being chosen for you by social media platforms? This question is becoming increasingly relevant as our default social media feeds shift from chronologically ordered content from sources (i.e., friends and followings) explicitly selected by us to algorithmically ordered content from any source implicitly inferred to match our interests by recommender systems. These systems aim to curate, i.e., select and order, personalized social media feeds based on users' data (either explicitly provided or implicitly inferred), such as their demographic, behavioral, and interest information. Algorithmically curated feeds are credited with increasing users' engagement with content and time spent on social media platforms, resulting in greater revenues for the platforms. However, concerns have been raised that algorithmic content curation is opaque and non-transparent to end users, takes choice and control over what content they see away from end users, and leads to undesirable phenomena, such as recommending content with extreme and polarizing viewpoints (TikTok 2020; Ribeiro et al. 2020; Ledwich and Zaitsev 2019; Tufekci 2018). As a result, there is a growing demand and need for greater transparency in social media feeds, particularly in how these recommendation systems operate, what user data they take into account, and how they impact users.

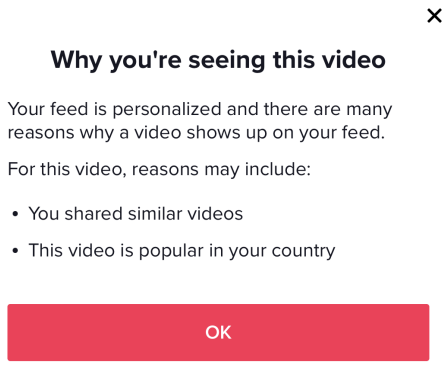


Figure 1: Example of video explanations on TikTok. Typically, a TikTok video has multiple *potential* explanations. In this example, the video has two potential explanations, as denoted by the bullet points.

Recently, regulatory bodies like the European Commission started working towards providing guidelines and regulations to assist algorithmic transparency efforts. Notably, the EU passed the Digital Services Act (European Commission 2023), which highlights the importance of algorithmic transparency and calls for systematic audits

of algorithmically-powered recommendation systems. The DSA attempts to hold very large social media platforms to account for systemic risks that may arise from the increasing use of algorithmically-powered recommendation systems. For instance, algorithmically-powered recommendation systems can trap end-users in filter bubbles (WSJ Staff 2021), suggest extreme content (Ribeiro et al. 2020), or amplify biases (Baeza-Yates 2020).

Against this background, social media platforms like TikTok started providing algorithmic explanations on why specific short-format videos are recommended to end-users in their feeds (TikTok 2022). Figure 1 shows an example of video explanations provided by TikTok for a recommended video in TikTok's "For You" feed. In this particular video, TikTok provided two explanations, each described by a bullet point. These efforts are a significant step towards greater algorithmic transparency. Particularly, accurate explanations can help end-users better understand why they view certain content on their personalized feeds and, if necessary, change or control the provision of such content. For example, recommendation systems can potentially push a user's feed towards a rabbit hole of sad content (WSJ Staff 2021). In this case, accurate explanations can help the end user to notice such a circumstance and take the appropriate action. Despite the importance of video explanations, as a research community, we lack 1) empirical insights and a deeper understanding of how these algorithmic explanations are applied; and 2) systematic audits to assess explanations' accuracy and comprehensiveness.

Research Focus. Motivated by this research gap, we focus on understanding video explanations on TikTok, mainly because of the increasing popularity of the platform and the fact that TikTok is among the few platforms that provide explanations for algorithmic recommendations. We aim to answer:

- **RQ1 - Characterization:** What are TikTok's various types of video explanations, and how does TikTok apply these explanations to videos?
- **RQ2 - Audit:** How accurate and comprehensive is the provision of video explanations on TikTok?

To answer these research questions, we designed and implemented a sockpuppet-based audit to collect a large-scale dataset of videos and explanations provided by the TikTok platform. We implemented sockpuppet accounts that do not engage with videos, as well as sockpuppet accounts that have engagement by either liking/sharing videos or following TikTok creators. We collected a dataset of 69K video views, including 32K unique videos, presented to 45 different sockpuppet accounts. Our sockpuppet accounts were presented with 128K explanations, comprising eight different explanations. Note that as shown in Figure 1, videos may contain multiple explanations. Next, we analyzed the collected dataset to understand how these TikTok explanations are applied. Finally, we operationalized multiple aspects of these explanations, such as their accuracy and comprehensiveness, and assessed their degree on TikTok.

Main findings. Our analysis makes the following findings:

- We find eight different explanations provided on TikTok videos and we broadly categorize them into engagement-based explanations (five explanations) and content-based explanations (three explanations) (**RQ1**).
- We find some explanations in almost all the videos; E.g., the explanation "This video is popular in your country" appears in over 98% of the videos in our dataset (**RQ1**).
- We find that explanations are written in a personalized manner but applied in a non-personalized way. For instance, we find the explanation "You commented on similar videos" to 34% of the videos that our sockpuppet accounts encountered when they made zero comments on the platform (**RQ1**).
- We find varying degrees of accuracy across the explanations. Overall, we find that content-based explanations are applied accurately on TikTok, however, when considering engagement-based explanations, we find a substantial number of false positives (i.e., there is an explanation, but we do not find any link with the account's behavior) or false negatives (i.e., there is no explanation but we do find a link with the account's behavior) (**RQ2**).
- We find that TikTok does not provide all possible explanations; for only 9% of the videos, TikTok provided all possible explanations (**RQ2**).

Implications. Our analysis and findings have important implications for various interested stakeholders, including social media operators, policymakers, and the research community. First, our analysis shows that the explanations are inaccurate and incomprehensive. All these aspects affect their utility to end-users, which highlights the need for social media platforms to develop better explainable algorithmically-based recommendation systems. For instance, when an explanation is presented in almost all the videos, it likely provides little helpful information to end-users, who will likely ignore these explanations. Our study is also of interest to policymakers who aim to assess the compliance of social media platforms with current regulations like the DSA. Our analysis shows that the provided explanations are likely not capturing user behavior effectively, indicating that these transparency efforts might not fully comply with the existing regulations. Finally, our study prompts further studies to assess the compliance of social media platforms with transparency regulations. The research community can build on top of our work to make longitudinal and multi-platform audits of algorithmic recommendations. To assist the research community in future research endeavors, we have made our code and dataset available, which can be used to perform longitudinal audits and further analysis of video explanations.[1]

2 Data Collection

This section describes our methodology for collecting our dataset. In a nutshell, we develop sockpuppet accounts that scroll through TikTok's "For You" page and capture the explanations provided to each video.

[1] https://github.com/sepehrmousavi/TikTok-Explanations

2.1 Sockpuppet Accounts

A substantial challenge when studying video explanations on TikTok is that the explanations are only available via TikTok's mobile applications, and there is no API that provides data about the explanations provided to users. To overcome this challenge, we develop a sockpuppet data collection methodology that uses the Android Debug Bridge (ADB) (Developers 2023) to control and simulate a user's behavior on an Android phone. Particularly, we created multiple TikTok accounts for our data collection. We registered all the sockpuppet accounts with the same date of birth and did not specify any gender for the accounts upon registration. For our data collection, a sockpuppet account first logins to the TikTok platform using one of the accounts we created. Then, the sockpuppet account visits TikTok's "For You" feed, and for each video, it: 1) identifies the type of video; 2) collects the explanations provided by TikTok; 3) engages with the video according to the sockpuppet's configuration; and 4) swipes up to the next video.

Identifying the type of videos. TikTok's "For You" feed includes different types of videos; apart from regular videos, it includes videos that are advertisements or live streams. Therefore, we must first identify the regular videos and skip videos that are advertisements or live streams, mainly because live streams do not have explanations, while advertisements have different explanations that require different steps to capture. To identify the type of a video, the sockpuppet account takes a screenshot and processes it using the OpenCV library (OpenCV 2023) to detect patterns pertaining to regular videos, advertisements, or live streams. Particularly, for live streams, we search for patterns in the screenshot that include the label "LIVE video," and do not include any buttons for engaging with the videos (the engagement buttons only appear after a user clicks on the video to go to the live stream). If the video is not a live stream, the sockpuppet account then presses the "Share" button, which opens the menu for the video; this allows us to identify if a video is an advertisement or not. This is because all advertisements should provide explanations, hence we search in the menu items for the button called "About this ad." If we find such a button, we mark the video as an advertisement, otherwise, we treat the video as a regular video.

Collecting video explanations. For all videos that we identified as regular videos, the sockpuppet account presses the "Share" button, takes a screenshot, and processes it to identify if there exists a "Why this video" button and locates it. If there is no such button, TikTok did not provide any explanation. If the button exists, the sockpuppet account presses the "Why this video" button, takes a screenshot and performs Optical Character Recognition (OCR) using Python-tesseract (Google 2023b) to extract the provided explanations in text format.

Engaging with videos based on configuration. After capturing the explanations, the sockpuppet account engages with the video based on the account's configuration. In this work, we consider three engagement activities: liking, sharing, and following. To like a video, the sockpuppet account taps on the heart icon. To share a video, the sockpuppet account taps on the "Share" panel and then taps on the "Copy link" button once the sharing panel pops up. [2] Lastly, to follow a content creator, the sockpuppet account first checks if the user is already followed or not by inspecting the screenshots of the video and finding the red plus icon located below the user's profile picture. If this icon is found, the sockpuppet account taps on it to follow the content creator.

Swiping up to the next video. Once the sockpuppet account engages with the current video, it swipes to view the next recommended content on its personalized "For You" feed. As shown by previous work, the view duration of a TikTok video influences the personalization of content for the end-users (Boeker and Urman 2022). To minimize this effect, the sockpuppet account swipes the screen up as soon as it identifies the type of video, collects the video explanations, and engages with the video. Our sockpuppet accounts typically swipe to the next video within 15 seconds (see Fig. 4(b)).

2.2 Experimental Setup

We designed and ran nine account configurations, each consisting of five sockpuppet accounts with identical behavior. Each sockpuppet account is controlled by a bot running on an Android mobile device and watches 1.8K videos divided into six sessions (300 videos in each session). Random delays separate each session to avoid overloading the TikTok platform. We ran five identical sockpuppet accounts for each configuration. Table 1 provides an overview of the configuration of our sockpuppet accounts.

No Engagement (NE) Configuration. The sockpuppet accounts with the NE configuration visit the "For You" feed, capture the explanations from all regular videos, and do not perform any engagement actions (no shares, no likes, etc.).

Liking (L1-L3) Configuration. These sockpuppet accounts visit the "For You" feed, and after capturing the explanations, they like videos with a probability of 0.05, 0.1, and 0.2 for L1, L2, and L3, respectively.

Sharing (S1-S3) Configuration. The sockpuppet accounts with the S1-S3 configurations are similar to the L1-L3 configurations, with the difference that instead of liking videos, they share a video with the same probabilities.

Following (F1-F2) Configuration. The sockpuppet accounts for the F1 and F2 configurations visit the "For You" feed, capture the video explanations, and follow the creator with a probability of 0.05 and 0.1 for F1 and F2, respectively.

During our data collection, we observed that TikTok drops many of the Like and Follow actions (see two last columns in Table 1); these drops are likely due to TikTok's anti-spam or anti-engagement-fraud mechanisms. These drops can potentially affect the recommended videos, however, since our work aims to study the video explanations, we believe these drops would not affect our main findings. Also, when performing our analysis, we consider only the actions recorded

[2] TikTok considers copying the video link as a share action as confirmed by obtaining data using GDPR access requests. Moreover, according to our investigation and empirical findings presented in Table 4, simply tapping on the "Share" panel is not considered as an engagement signal by TikTok.

Config	$P(Like)$	$P(Share)$	$P(Follow)$	#Bot Actions	#TikTok Actions
NE	0	0	0	0	0
L1	0.05	0	0	385	263
L2	0.10	0	0	776	418
L3	0.20	0	0	1,485	776
S1	0	0.05	0	367	366
S2	0	0.10	0	820	818
S3	0	0.20	0	1,465	1,464
F1	0	0	0.05	388	159
F2	0	0	0.10	741	252

Table 1: Configuration of our automated accounts. We report the probabilities that an account will like, share, and follow a video appearing on the "For You" feed. We also report the number of attempted actions performed by the account and the number of actions that TikTok recorded.

by TikTok and disregard not-recorded actions made by our sockpuppet accounts.

2.3 Data Collection

We ran our data collection between March 28, 2023, and May 5, 2023. For each sockpuppet account, we requested the account's activity using the right of access to data subjects described by the EU General Data Protection Regulation (Commision 2016), similarly to (Zannettou et al. 2023). This allows us to obtain the account's entire activity, including all the video identifiers that the user watched and all the actions recorded by the TikTok platform. This is a crucial and necessary step, as there is no easy way to obtain the video identifier from the TikTok mobile application.[3] Then, we match the explanations extracted from the screenshots and the videos based on the timestamps. Also, after obtaining the datasets from TikTok, following our GDPR request, we perform a metadata collection for each video using an unofficial Python wrapper for the TikTok API (Teather 2022). For each video, we obtain the video's title, description, and associated metrics, such as the number of views, likes, shares, and comments that the video received on TikTok. We perform this additional metadata collection, as the data provided by TikTok includes only the video identifiers without any metadata about each video. Overall, our sockpuppet accounts viewed 81,000 videos; 70,065 regular videos, 10,132 advertisement videos, and 803 live streams. In this work, we focus on regular videos, and after excluding the videos with no metadata (less than 1%), our dataset consists of 69,430 video views, including 32,553 unique videos. Our sockpuppet accounts identified explanations in 63,394 video views, and TikTok provided a total of 128,324 video explanations (eight unique explanations, see Table 2).

3 RQ1: Characterizing Video Explanations

Here we present our characterization of video explanations. We discuss the various types of explanations in Section 3.1, while Section 3.2 presents our analysis for characterizing how TikTok applies the explanations, how prevalent each type of explanation is, and how the explanations are correlated with user engagement and video characteristics.

[3]The only way is to press the "Share" button and copy the video link. However, this is considered a share action, and we want to avoid doing this as this will mean that our accounts share all videos.

Category	Explanation
Engagement-based	You **shared** similar videos
	You **commented** on similar videos
	You **liked** similar videos
	You are **following** <username>
	You **watched** similar videos
Content-based	This video is **longer**, and you seem to like longer videos
	This video is **popular** in your country
	This video was posted **recently**

Table 2: List of video explanations. The bold text refers to the name of each explanation (used throughout the paper).

3.1 Types of Video Explanations

We start by looking at the explanations that we discovered (see Table 2). Our sockpuppet accounts encountered eight explanations, and we broadly categorize them into *Engagement-based Explanations* and *Content-based Explanations*.

Engagement-based explanations consider user engagement when generating the explanations. Our sockpuppet accounts encountered five Engagement-based explanations:

- **Shared:** "You shared similar videos." Refers to videos similar to the videos a user has previously shared on TikTok or other platforms (e.g., WhatsApp, Facebook, etc.).
- **Commented:** "You commented on similar videos." Refers to videos that are similar to the videos that a user commented on.
- **Liked:** "You liked similar videos." Refers to videos that are similar to the videos a user previously liked.
- **Following:** "You are following $< username >$." Refers to videos uploaded by a creator that a user has already followed. This is a parameterized explanation, as it pro-

Config.	#Videos	%Ad Videos	#Reg. Videos	%Reg. Videos w. Explanations
NE	8,943	13.98 %	7,645	92.35 %
L1	8,912	11.74 %	7,804	87.03 %
L2	8,959	10.36 %	7,952	89.96 %
L3	8,927	12.98 %	7,684	94.47 %
S1	8,893	11.02 %	7,818	97.93 %
S2	8,917	12.37 %	7,728	96.27 %
S3	8,958	15.73 %	7,444	98.27 %
F1	8,917	11.73 %	7,731	88.42 %
F2	8,939	13.55 %	7,624	77.15 %

Table 3: Video statistics for each configuration.

vides a creator's username at the end.
- **Watched:** "You watched similar videos." Refers to videos that are similar to the previously viewed videos that a user watched until the end.

Content-based explanations are provided because of content characteristics, such as the content length, the popularity of the content, and when it was posted. Our sockpuppet accounts encountered three Content-based explanations;
- **Longer:** "The video is longer, and you seem to like longer videos." Refers to videos that have a long duration. Note that the phrasing of this explanation hints toward both Engagement and Content signals. However, based on our analysis in Section 3.2, we argue that it is a primarily Content-based explanation.
- **Popular:** "This video is popular in your country." Refers to videos that are popular in the user's country.
- **Recently:** "The video was posted recently." Refers to videos recommended to a user shortly after their creation.

Overall, by reading these explanations, it becomes apparent that several aspects of these explanations are vague, which can potentially limit their utility to end-users. For instance, many explanations include "similar videos," without providing context on assessing the content similarity. Also, there are no clear guidelines on what constitutes a long, popular, or recent video when providing these explanations.

3.2 Prevalence and Characterization of Explanations

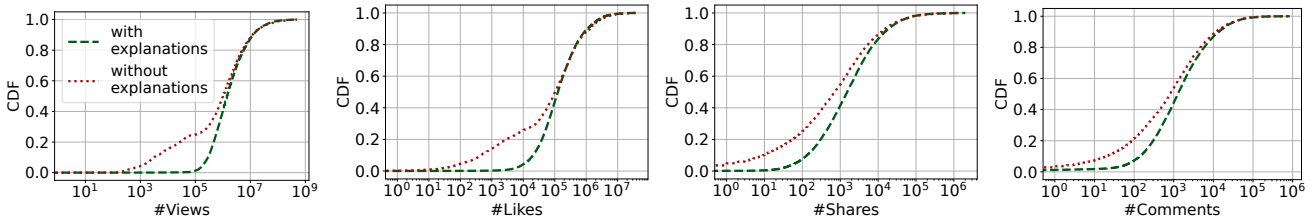
Here, we characterize and measure the prevalence of video explanations. Table 3 reports the number of videos watched by our sockpuppet accounts for each account configuration and the percentage of videos with video explanations. We observe that most videos have explanations, on average, 91.31% of all videos. Also, by examining the number of explanations per video, we find that 8.69% of the videos have no explanations, 89.10% of the videos have two explanations, 2.18% of the videos have three explanations, while 0.02% of the videos have four explanations. Our sockpuppet accounts encountered 10%-15% of videos that were advertisements; all the advertisements had explanations that were completely different from those provided in regular videos. Given that ad explanations are extensively studied by previous work (Andreou et al. 2018; Lee et al. 2022; Wei et al. 2020; Wilkinson et al. 2021; Gkiouzepi et al. 2023), in our analysis, we focus on explanations provided to content that is recommended by the TikTok algorithm and is not advertisements, namely regular videos.

Prevalence of video explanations. Next, we aim to measure and analyze the prevalence of each explanation. We do this analysis to potentially understand how TikTok applies these explanations and explore the factuality of video explanations. Table 4 reports, for each configuration, the percentage of videos that had each video explanation (each row reports the mean and standard deviation of five sockpuppet accounts). We make several observations. For the NE configuration, where the sockpuppet accounts scroll through the "For You" feed without engaging with the videos (other than watching the videos for a few seconds), we observe a surprisingly large percentage of Engagement-based explanations. Specifically, 34% of the videos received the *Commented* explanation, 8% the *Liked* explanation, and 0.67% the *Shared* explanation. These results are surprising, given that the sockpuppet accounts did not comment, like, or share any video in this account configuration. Second, we observe that the *Popular* explanation is provided to almost all videos across all configurations (more than 98%). Third, for the configurations involving user engagement (i.e., liking, sharing, and following accounts), we observe that the prevalence of the explanations changes based on the sockpuppet behavior. For instance, when the sockpuppet accounts like videos (see configurations L1-L3), the *Liked* explanation is provided to 89%-97% of all videos; however, we still observe videos receiving explanations for other engagements that the account did not perform (e.g., 2%-6% for *Commented* explanation). The same finding applies when sockpuppet accounts share videos (see configurations S1 to S3) or followed accounts (configurations F1 and F2). For the *Following* explanation, we observe that it is only provided to the sockpuppet accounts that followed some accounts with only a small percentage of videos (3%-6%). Finally, we observe that the *Recently* explanation is not frequent; we only encountered it in 0.03% of all videos in the S1 configuration.

Videos with explanations vs. without explanations. Here, we aim to uncover potential underlying reasons why a considerable percentage of videos (8.69%) do not have explanations. Given that almost all videos have the *Popular* explanation, a potential reason for not including explanations for some videos is that those videos are not popular on TikTok. To assess if this is a valid reason, we compare the distributions of various popularity metrics (number of views, likes, shares, and comments) for videos that received explanations and videos without explanations. We use these metrics as a proxy for content popularity on TikTok, which is in line with previous work (Klug et al. 2021; Boeker and Urman 2022). Figure 2 shows the Cumulative Distribution Function (CDF) of the number of views, likes, shares, and comments for videos with explanations and videos without explanations for all the videos in our dataset. We observe that videos with explanations generally have more views, likes, shares, and comments than videos without explanations. No-

Config	%Shared	%Commented	%Liked	%Following	%Watched	%Longer	%Popular	%Recently
NE	0.67 ± 0.5	34.09 ± 5.5	8.01 ± 0.7	0.00 ± 0.0	56.00 ± 3.7	1.93 ± 2.2	100.00 ± 0.0	0.00 ± 0.0
L1	0.03 ± 0.0	6.78 ± 2.8	89.62 ± 3.9	0.00 ± 0.0	3.51 ± 3.3	5.52 ± 6.7	99.97 ± 0.0	0.00 ± 0.0
L2	0.00 ± 0.0	2.71 ± 1.8	97.24 ± 1.8	0.00 ± 0.0	0.01 ± 0.0	1.28 ± 1.5	99.99 ± 0.0	0.00 ± 0.0
L3	0.00 ± 0.0	4.22 ± 1.1	95.78 ± 1.1	0.00 ± 0.0	0.00 ± 0.0	2.04 ± 2.7	99.97 ± 0.1	0.00 ± 0.0
S1	94.95 ± 1.5	3.70 ± 0.9	0.83 ± 0.5	0.00 ± 0.0	0.48 ± 0.9	1.90 ± 2.0	99.95 ± 0.0	0.03 ± 0.1
S2	97.84 ± 0.9	1.68 ± 0.7	0.48 ± 0.3	0.00 ± 0.0	0.00 ± 0.0	0.55 ± 0.4	99.99 ± 0.0	0.00 ± 0.0
S3	98.46 ± 0.4	1.27 ± 0.3	0.26 ± 0.2	0.00 ± 0.0	0.00 ± 0.0	0.82 ± 0.5	99.99 ± 0.0	0.00 ± 0.0
F1	0.07 ± 0.1	31.63 ± 15.9	14.06 ± 13.2	2.99 ± 1.3	51.98 ± 12.7	3.66 ± 3.1	99.66 ± 0.3	0.00 ± 0.0
F2	0.07 ± 0.1	59.65 ± 13.1	20.52 ± 10.3	5.86 ± 3.2	19.40 ± 22.7	2.31 ± 2.2	98.54 ± 1.2	0.00 ± 0.0

Table 4: Percentage of videos that have each explanation. We focus on the set of videos that had at least one explanation.Each row contains, for each configuration, the mean and standard deviation of the percentage of videos that had the corresponding explanation across the five sockpuppet accounts.



(a) Views (b) Likes (c) Shares (d) Comments

Figure 2: CDF of the number of views, likes, shares, and comments on TikTok for videos with and without explanations.

tably, for views (see Figure 2(a)), we observe that almost all the videos with explanations have more than 100K views, whereas, for videos without explanations, we observe a considerable percentage of videos that are below 100K views (25%). The same applies to the likes (see Figure 2(b)); most of the videos with explanations have more than one thousand likes, while 14% of the videos without explanations have less than one thousand likes. For shares and comments, we cannot find a clear cut-off lower-bound threshold for videos with explanations; still, the videos with explanations tend to have more shares and comments than videos without explanations.

Video Explanations and Video Engagement. Here, we investigate if there is any correlation between the video metadata and the provided explanations. Our results, thus far, show that accounts that do not engage with videos on TikTok are provided with a substantial number of engagement-based explanations (i.e., liking, sharing, and commenting). One question that arises is whether TikTok provides such explanations to videos that generally have very high engagement on the entire platform, irrespective of the user's actions. In other words, we aim to investigate, for instance, if TikTok is more likely to provide the *Liked* explanation to videos that have already accumulated a lot of likes on TikTok. Figure 3 shows the CDF of the number of likes, shares, and comments for videos that had the *Liked*, *Shared*, and *Commented* explanation and videos without those explanations, respectively, for the NE configuration. We observe that videos that receive the *Liked* explanation have substantially more likes on TikTok compared to the videos without the *Liked* explanation. The same applies to the videos that receive the *Shared* and *Commented* explanations. The differences in these distributions are significant, as confirmed by a two-sample Kolmogorov-Smirnov test ($p < 0.01$). We repeated the same analysis for the rest of the configurations, finding similar statistically significant results. Overall, these results imply that the engagement received by a video on the entire TikTok platform will likely affect the provided explanations, as we observed that even for accounts with no engagement, we have videos that receive these explanations, likely because they already have a lot of engagement.

Video Explanations and Video Characteristics. We also compare the video explanations with video-based characteristics like the video duration. Particularly, we compare the distributions of the video duration for videos that received the *Longer* explanation and videos without the *Longer* explanation (see Figure 4(a)). We observe that all videos that receive the *Longer* explanation have a duration of 45 seconds or more, whereas the overwhelming majority of videos without the *Longer* explanation have a duration of less than 45 seconds. Recall, however, that the explanation is phrased as "This video is longer, and you seem to like longer videos," which indicates that TikTok also considers the user behavior when providing this explanation. To investigate whether TikTok actually considers user engagement (in the form of watching the videos till the end), we plot the CDF of how much time our sockpuppet accounts spend on each video to capture the explanations (see Figure 4(b)). We observe that

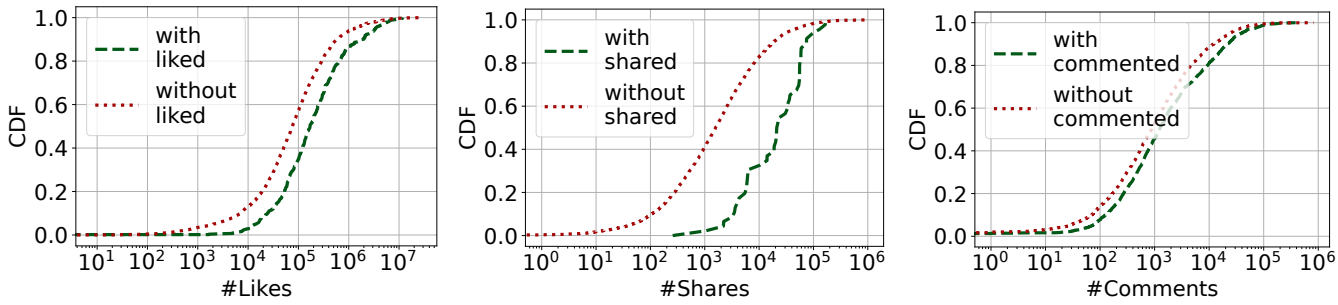


Figure 3: Comparison of video metadata for videos with/without an explanation for the No-Engagement (NE) configuration.

for 92% of the videos, our sockpuppet accounts spent 15 seconds or less on them, with a maximum value of 32 seconds. This indicates that our sockpuppet accounts did not watch till the end any of the "long" videos (i.e., 45 seconds or more), which likely indicates that the explanation is solely provided because the video has a long duration, irrespective of how long the user watched previous so-called "long" videos, which is a proxy to identify if a user likes the content.

3.3 Main Take-Aways

The main take-away points from our analysis aiming to characterize video explanations are:

- A large percentage of videos (more than 40%) received engagement-based explanations when the sockpuppet accounts did not engage with any video.
- We find that the *Popular* explanation is provided to almost all the videos that receive explanations on TikTok (over 98% across all our sockpuppet accounts).
- TikTok fails to provide explanations to 8.69% of all the video recommendations.
- The use of engagement-based explanations is likely influenced by the overall engagement that videos received on the entire TikTok platform. For instance, we find that videos with the *Liked* explanation have significantly more likes than videos without the *Liked* explanation.
- The TikTok explanations are written in a way that suggests they are personalized based on the activity/engagement of a user, however, our analysis indicates that explanations are applied in a non-personalized way (i.e., explanations are likely added due to the overall video engagement/characteristics, irrespective of users' activity/engagement).

4 RQ2: Accuracy and Comprehensiveness of Video Explanations

Thus far, our analysis shows that some explanations are provided to videos in a way that does not consider user behavior. Motivated by these preliminary findings, in this section, we take a deep dive into understanding two key aspects of video explanations, namely Accuracy and Comprehensiveness. To do this, we create connections between the user behavior/video characteristics and the respective explanations,

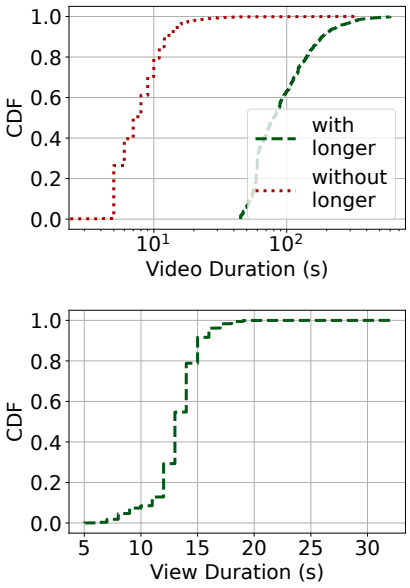


Figure 4: CDF of (a) video durations for videos with/without the *Longer* explanation and (b) the sockpuppet accounts' view duration for each video.

hence labeling the videos with TikTok explanations. For accuracy, we model the problem as a binary classification task and measure the number of false positives and false negatives generated between our labeling and TikTok-provided explanations. For comprehensiveness, we compare the set of our labeled explanations and those provided by TikTok to assess if TikTok provides all the applicable explanations on videos. Below, we describe our methodology for operationalizing these aspects and our findings.

4.1 Methodology and Data Preprocessing

Labeling Methodology. To assess the accuracy and comprehensiveness of video explanations, we need to label the videos with all feasible explanations based on user behavior and video characteristics. By doing so, we generate a ground truth dataset of explanations, which we will later compare with the TikTok-provided explanations to assess the accu-

racy and comprehensiveness of the explanations. Our labeling approach relies on the results presented in Section 3.2 with some additional assumptions on what constitutes "similar videos," which is a phrase commonly used in TikTok explanations (see Table 2). Our labeling approach works by iterating over the videos chronologically and assigning all possible explanations to each video based on some pre-determined conditions for the engagement-based and content-based explanations. We expand over these underlying "assumptions" and "pre-determined conditions" that were used in our labeling process in the paragraphs.

The four Engagement-based explanations, namely *Shared*, *Liked*, *Watched*, and *Commented*, express the "similarity" between videos as a possible reason for a video being recommended. Therefore, we need to measure the similarity between the videos so that our explanation labeling approach can assign these four explanations to the videos. To measure the similarity between videos, we consider their hashtags, sounds, and content creators; we assume two videos are similar if they share at least one common hashtag, or have the same sound, or are created by the same content creator. We do not provide any of these four explanations in cases where videos do not have hashtags or a sound. Given this notion of video similarity, we assign these four explanations to a video taking into account the account's engagements with previous videos on its feed. Particularly, a video gets the *Shared*, *Liked*, or *Commented* explanations if there exists a similar video on its feed that was previously shared, liked, or commented on. As for the *Watched* explanation, we consider a video as watched if the account watched it for at least the full duration of the video. For the *Following* explanation, we check if a video's content creator was already followed.

For the Content-based explanations, our labeling approach relies solely on video characteristics. For the *Longer* explanation, we consider a video to be long if its duration is at least 45 seconds; we select this threshold based on the analysis in Section 3.2 (see Figure 4(a)), as we aim to have a labeling approach that is as much as possible inline with how TikTok adds explanations. Similarly, for the *Popular* explanation, we assign the explanation if a video has at least 100K views (see Figure 2(a)). Finally, for the *Recently* explanation, our dataset contains only two videos with this explanation. These instances have a time difference between the video view and upload item of 81K and 77K seconds. Since we lack sufficient data to draw conclusions as to what constitutes a "recent" video, we refrain from providing the *Recently* explanation to the videos.

Preprocessing the Dataset. One of the features used to assess video similarity is the set of hashtags provided by the video creator in the video description. On TikTok, many content creators include various generic and popular hashtags to influence the recommendation system and increase the probability of the video appearing in other users' "For You" feed (Klug et al. 2021). For example, such generic and popular hashtags are #foryoupage, #fyp (an acronym for "For You Page"), #foryou, and #viral. The existence of these hashtags is affecting our similarity approach; i.e., if all videos have the #fyp hashtag, then our labeling approach will consider all videos as similar to each other. Therefore, we preprocess our dataset and remove all generic hashtags from the video descriptions. To do this, we consider the hashtags that appear in at least 0.1% of the videos, corresponding to the top 470 hashtags in our dataset. Next, we review these hashtags systematically and remove the general and meaningless ones. These removed hashtags are generally different variations of writing the #foryoupage or #viral hashtags.

Explanation	FP	FN	TP	TN	NA
Shared	0.12	0.00	0.10	0.68	0.10
Commented	0.14	0.00	0.00	0.86	0.00
Liked	0.17	0.01	0.08	0.65	0.09
Following	0.00	0.00	0.01	0.99	0.00
Watched	0.02	0.45	0.08	0.17	0.28
Longer	0.00	0.00	0.02	0.98	0.00
Popular	0.01	0.07	0.90	0.02	0.00

Table 5: Accuracy metrics for each explanation provided by TikTok. We report the False Positive rate (FP), the False Negative rate (FN), the True Positive rate (TP), and the True Negative rate (TN). NA refers to the fraction of videos for which we cannot assess their similarity (videos with no hashtags or sound). Note that as discussed in Section 4.1, we do not consider the *Recently* explanation in our accuracy analysis.

Validating the Labeling Methodology. Our labeling methodology makes some assumptions about what constitutes similar videos, and we do not know how accurate our similarity labels are. To assess how accurate our methodology is concerning labeling similar videos, we perform a manual validation. We extract a random sample of 100 video pairs: 50 video pairs labeled as similar and 50 video pairs labeled as dissimilar. Then, three researchers independently annotated the video pairs to assess whether they were similar. Given that TikTok does not provide a notion of similarity, we asked the annotators to label the videos as similar or dissimilar without providing a definition of similarity. We find a Fleiss' Kappa score of 0.825, which indicates an almost perfect inter-annotator agreement between the three researchers (Fleiss 1971). In the end, we label the 100 videos in our sample based on the majority agreement of the labels and calculate standard classification metrics to assess the performance of the labeling methodology. We observe that our similarity labels have an accuracy, precision, recall, and F1 score of 0.84, 0.78, 0.88, and 0.82, respectively. We believe that for the purposes of this study, this performance is acceptable.

4.2 Results

Here we present our results on the accuracy and comprehensiveness of TikTok explanations.

Accuracy. By considering the provision of an explanation to a video as a binary classification task, we can assess its accuracy by measuring the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) metrics.

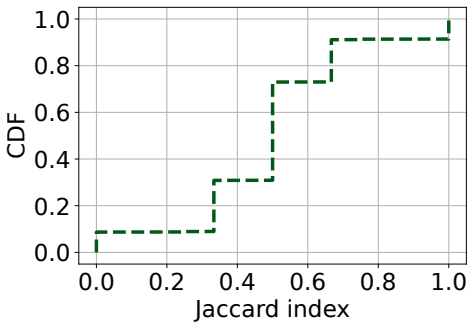


Figure 5: CDF of the Jaccard index between the explanations generated by our methodology and TikTok explanations.

Here, we treat the explanations provided by TikTok as predictions and the explanations provided by our labeling approach as ground truth labels. As noted in Section 4.1, for the explanations that require assessing video similarity, we are unable to assess the similarity for videos without hashtags or sound, hence, we put them in the Not Available (NA) category.

Table 5 shows the aggregate accuracy metrics of each explanation across all account configurations. The *Shared* explanation has a FP rate of 0.12, FN of 0.0, TP of 0.10, TN of 0.68, and NA of 0.10. Therefore, we see a moderate total error rate of 0.12 and a high accuracy of 0.78. For the *Commented* explanation, we find a FP of 0.14 and TN of 0.86. The rest of the metrics are 0. As no account engaged with TikTok by writing comments on the videos, any occurrence of the *Commented* explanation would be a False Positive instance. Hence, these results are indeed in accordance with our experimental setup. Our findings show a FP rate of 0.17, FN of 0.01, TP of 0.08, TN of 0.65, and NA of 0.09 for the *Liked* explanation. With a moderate total error rate of 0.18 and an accuracy rate of 0.73, we note that TikTok's recommendation system acts quite similarly in terms of accuracy for both the *Liked* and *Shared* explanations. We find a 0.01 rate of TP and a 0.99 rate of TN for the *Following* explanation, while the other metrics are 0. Our results show that this explanation is one of the most accurate explanations provided by TikTok. For the *Watched* explanation, we observe a FP rate of 0.02, FN of 0.45, TP of 0.08, TN of 0.17, and NA of 0.28. Our findings show that with an error rate of 0.47, the *Watched* explanation is the most inaccurate among all seven explanations. The *Longer* explanation has a TP rate of 0.02, a TN of 0.98, and the rest of the metrics are at 0. Similar to the *Following* explanation, our results indicate that this explanation, while being almost error-free, is also one of the most accurate explanations provided by TikTok. Lastly, for the *Popular* explanation, we see a FP rate of 0.01, FN of 0.07, TP of 0.9, TN of 0.02, and NA of 0. With an error rate of 0.08, this explanation is also among the most accurate explanations provided by TikTok.

Comprehensiveness. By treating the explanations provided by TikTok and our labeling explanations for a video as two sets, we can assess the comprehensiveness of the provided explanations by measuring the Jaccard index of these two sets. Figure 5 shows the CDF of the Jaccard index for all the videos in our dataset. We observe that in 9% of the videos, we have a Jaccard index of 0, implying a complete disagreement between the set of explanations provided by TikTok and our labeled explanations. Moreover, in 9% of the videos, we have a Jaccard index of 1, indicating full agreement between the set of explanations provided by TikTok and our explanation generation system. Also, we see that in 69% of the videos, the computed Jaccard index is at least 0.5, suggesting that a fair amount of comprehensiveness exists in the explanations provided by TikTok. Overall, these results suggest that for some videos, there are possible explanations that are not provided by TikTok, which highlights the need for more comprehensive transparency.

4.3 Main Take-Aways

The main take-away points from our analysis to assess the accuracy and comprehensiveness of video explanations are:

- Our accuracy assessment analysis shows that overall, the explanations *Following* and *Longer* are the most accurate explanations provided by TikTok's recommendation system. Followed by these, the *Popular* and *Commented* explanations, with error rates of 0.08 and 0.14, respectively, obtain the highest accuracies among the remaining explanations. On the other hand, the *Shared*, *Liked*, and *Watched* explanations have lower accuracies, with the *Watched* being the least accurate explanation with an error rate of 0.47, and *Liked* and *Shared* explanations having error rates of 0.18 and 0.12, respectively.
- Our comprehensiveness analysis shows that TikTok does not provide all the possible explanations for video recommendations. We find that for only 9% of the video recommendations, we had the same explanations provided by TikTok and our labeling approach.

5 Related Work

This section describes previous efforts to understand content explanations and auditing recommendation systems.

Explanations for Social Media Content. The prior work on the explanations provided by social media platforms as to why content is being suggested to the users is mainly focused on the explanations of the advertisements. This is because, until recently, mainstream social media platforms had not incorporated the feature of providing explanations for normal content other than advertisements. As a groundbreaking step towards transparency and explainability of the contents shown on users' feeds, the giant social media platforms started providing explanations for advertisements, shedding light as to why certain ads are shown to the users (Meta 2023; Google 2023a; Corp. 2023). Back in 2018, in a seminal piece of work, Andreou et al. (Andreou et al. 2018) performed the first empirical study of explanations in social media advertising and showed that in the case of Facebook, the provided ad explanations are usually incomplete and even sometimes misleading. With the increasing number of machine learning applications, algorithmically-based recommendation systems

are also emerging. Prior work has stated the importance of defining and evaluating interpretability, transparency, and accountability of these machine learning systems (Lipton 2018; Doshi-Velez and Kim 2017; Wieringa 2020; Weller 2017). On the other hand, with the advent of these machine learning models and their prevailing acceptance by society, the concept of *trust* in these systems, from the perspective of the users, is becoming more critical. Some previous work indicates that providing explanations and adopting transparency would increase public trust in a platform (Lipton 2018; Ribeiro, Singh, and Guestrin 2016). Despite this, Weller (Weller 2017) argues that platforms may be able to trick their users into having trust in their systems by providing useless explanations. On the other hand, Kizilcec (Kizilcec 2016) shows that providing either too little or too much transparency to the users erodes their trust in the systems. Therefore, platforms require balanced interface transparency, not too little and not too much, when aiming to maximize the users' trust. Consequently, finding this balanced sweet spot is a challenging task. In work towards this direction focused on ad explanations, researchers found that the users preferred interpretable and non-creepy explanations, as well as a recognizable link to their identity as to why an ad is shown to them (Eslami et al. 2018).

Recommendation Systems Auditing. With the increasing advancement and utilization of machine learning systems, user feeds on social media platforms have started serving content that is suggested by the recommendation algorithms (TikTok 2020). Therefore, the importance of these recommendation systems and their effects on humans are growing over time (Sandvig et al. 2014; Bandy 2021). On the other hand, providing detailed explanations on how these neural network-based systems work has been a challenging task (Mitchell et al. 2007; Goodfellow, Bengio, and Courville 2016). Hence, to better understand how these artificially intelligent recommendation systems work, we need to conduct audits (European Commission 2023). Researchers have already shown that algorithm audits can reveal controversial and problematic behaviors of recommendation systems. For instance, recommendation systems may trap their users in filter bubbles (WSJ Staff 2021), provide them with extreme contents (Ribeiro et al. 2020), or preserve biases (Baeza-Yates 2020). To audit the video explanations provided by TikTok platform, we created fake user profiles and developed an automated bot that engages with the videos on their "For You" feeds. According to the characterization of audit study designs by (Sandvig et al. 2014), our approach belongs to the sockpuppet audit category. There have been multiple pieces of work that are based on sockpuppet audits. For instance, an investigation on TikTok's recommendation system based on sockpuppet audits showed that TikTok may be able to quickly figure out the bot's interests, or push the bot's feed towards a rabbit hole of sad videos (WSJ Staff 2021). In another work, Boeker and Urman empirically investigated various personalization factors on the TikTok platform using a sockpuppet audit methodology (Boeker and Urman 2022).

Remarks. In contrast to previous work, we focus on auditing video explanations of non-advertisement content on TikTok. To the best of our knowledge, we perform the first systematic audit of explanations provided by social media platforms on online content recommended by algorithms and not sponsored/paid. We believe that our work paves the way toward auditing transparency efforts by very large social media platforms like TikTok, by characterizing these explanations and assessing their accuracy and comprehensiveness.

6 Ethical Considerations and Broader Impact

In this section, we discuss our ethical considerations when conducting this research and the FAIR principles for releasing our dataset. Our study focuses exclusively on collecting publicly available information on the TikTok platform using sockpuppet accounts, hence we are not dealing with any sensitive user data. Our sockpuppet accounts performed some actions on TikTok (i.e., like videos and randomly follow some creators), however, since we only used 45 sockpuppet accounts in total, we do not anticipate that any user harm will arise from these actions. At the same time, we carefully designed our sockpuppet accounts to not incur a substantial overhead to the TikTok platform (e.g., by adding delays between sessions), which might harm the user experience of other TikTok users browsing the platform or cause any disruption to the platform. Overall, even though TikTok's Terms of Service (ToS) prohibits any data or content extraction from the platform using any automated system or software that is not provided by TikTok or approved in writing by TikTok (TikTok 2023), we believe that the benefits of our systematic audit and analysis outweigh the potential harms arising from our data collection and interaction with the TikTok platform.

We plan to make our dataset publicly available following the FAIR principles (Wilkinson et al. 2016). The dataset will be available on a prominent cloud storage service, making it *findable* and *accessible*. The dataset will be *interoperable* in that it will be released in a Python-readable format. Moreover, our dataset is *reusable* as besides clearly stating the steps taken to collect the dataset, we plan to release a README file that explains correct data usage.

7 Discussion & Conclusion

In this work, we performed a large-scale audit of explanations provided on TikTok video recommendations. Using 45 sockpuppet accounts, we collected 69K video views and 128K explanations. We plan to make the code and collected dataset publicly available, which we believe is an essential step towards more algorithmic audits and understanding these transparency efforts. Then, we analyzed the collected dataset to understand the various video explanations and how they are applied on TikTok. Also, we assessed the accuracy and comprehensiveness of these explanations on TikTok. Our characterization shows that TikTok applies several Engagement-based and Content-based explanations on each video recommendation, with almost 90% of the videos including two video explanations. Also, we find that some explanations, e.g., the *Popular* explanation, are provided in

almost all the videos (more than 98% of the videos with explanations). More worrying is the finding that some explanations are written in a personalized manner, but they are applied in a non-personalized manner; e.g., we find that 34% of the videos include the *Commented* explanation when our sockpuppet account made zero comments on TikTok. Finally, our accuracy analysis indicates that TikTok explanations have varying accuracies, with content-based explanations being more accurate than engagement-based explanations. Concerning comprehensiveness, we find for most of the videos (91%), TikTok does not provide all feasible explanations. Below, we discuss the implications of these findings and the limitations of this work.

Explanation Utility. Our work emphasizes the need for further studies to investigate the utility of video explanations. We observed that explanations are sometimes added without considering the user behavior and that some explanations appear in almost all the videos. These two factors affect the utility of the explanations for end-users. For instance, end-users will likely ignore explanations that are presented in all videos, as they are not informative, while the fact that explanations are added without considering the user behavior may cause end users not to trust the provided explanations (i.e., the perceived explanation accuracy will diminish).

Explanation Personalization & Phrasing. Our characterization shows that most explanations are written and presented in a way that they are personalized, however, our analysis shows that they are applied in a non-personalized way. This highlights the need for greater transparency by social media platforms, particularly explaining to users which explanations pertain to their user engagement/behavior and which explanations are not. We argue that it is important that social media platforms design more fine-grained explanations that will give users enough context to understand why they are getting recommended certain videos.

Regulation Compliance. DSA's recital 70 mentions that online platforms "should include at least the most important criteria in determining the information suggested to the recipient of the service and the reasons for their respective importance, including where information is prioritised based on profiling and their online behaviour" (European Commission 2023). Based on our analysis and findings, we can conclude that TikTok's transparency efforts are not fully compliant with the DSA regulation, given that most of the provided explanations are generic and there is no explanation of the importance of the factors that affect the recommendations. Taken all together, there is a need for collaboration between social media platforms, regulators, and end-users to ensure that the social media platforms are providing explanations that are compliant with the regulations and that end-users can comprehend why they are getting recommended specific content online easily.

Potential Sources of Bias. Our analysis and findings demonstrate that TikTok video explanations on regular videos are very generic and high-level. Therefore, it is highly unlikely that these explanations can be used to analyze potential sources of bias. For instance, whether TikTok uses explanations in a specific way in order to fit some political agenda. Nevertheless, some of our preliminary investigations on explanations in advertisement videos show that those explanations are more fine-grained and specific. For instance, we find explanations like "TikTok's estimate of your interest in News & Entertainment or a similar category" or "TikTok's estimate of your interest in Sports & Outdoors or a similar category." This indicates that TikTok has more information on the users, which is not provided in the explanations on regular videos. We believe that in the future, a potential research direction is analyzing the potential sources of bias by looking into explanations provided on both regular and advertisement videos.

Limitations. Our work has some limitations. First, our data collection is limited to a single platform (i.e., TikTok) and it is done for a short time period (3 months). These limitations do not allow us to study the use of video explanations across multiple platforms, or how the time dimension may affect the provided video explanations. We leave all these research endeavors as part of our future work. Second, our accuracy and comprehensiveness analysis relies on our labeling approach that assesses the similarity of videos using hashtags, sounds, and video creators. Based on the user annotation results for validating our labeling approach, we think that these features are a good proxy for content similarity, however, our approach is limited because even though the features we use can be potential indicators of video similarity, we do not consider the complex nature of video content. Despite these important limitations, we believe that our work is an important step towards understanding the application and effectiveness of these algorithmic transparency efforts by very large social media platforms like TikTok.

Future Work. Our work creates avenues for future work. Future work can explore the causal relationships between the quality and personalization of explanations and user behavior, trust, and content engagement. Also, future work can investigate differences in video explanations across various demographics, as well as investigate ways to study biases that may arise from the use of video explanations by platforms like TikTok.

References

Andreou, A.; Venkatadri, G.; Goga, O.; Gummadi, K. P.; Loiseau, P.; and Mislove, A. 2018. Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *NDSS*, 1–15.

Baeza-Yates, R. 2020. Bias in search and recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 2–2.

Bandy, J. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *CSCW*, 5: 1–34.

Boeker, M.; and Urman, A. 2022. An Empirical Investigation of Personalization Factors on TikTok. In *Proceedings of the ACM Web Conference 2022*, 2298–2309.

Commision, E. 2016. General Data Protection Regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed: 2023-05-20.

Corp., X. 2023. Why are you seeing this ad . https://help.twitter.com/en/why-are-you-seeing-this-ad. Accessed: 2023-05-20.

Developers, A. 2023. Android Debug Bridge (adb). https://developer.android.com/tools/adb. Accessed: 2023-05-20.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Eslami, M.; Krishna Kumaran, S. R.; Sandvig, C.; and Karahalios, K. 2018. Communicating algorithmic process in online behavioral advertising. In *CHI*, 1–13.

European Commission. 2023. The Digital Services Act package. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package. Accessed: 2023-05-20.

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

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2023-09-10.

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.

Gkiouzepi, E.; Andreou, A.; Goga, O.; and Loiseau, P. 2023. Collaborative Ad Transparency: Promises and Limitations. In *44th IEEE Symposium on Security and Privacy*.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Google. 2023a. Control the ads you see when you see them . https://support.google.com/My-Ad-Center-Help/answer/12155764. Accessed: 2023-05-20.

Google. 2023b. Python Wrapper for Google's Tesseract OCR Engine. https://pypi.org/project/pytesseract/. Accessed: 2023-05-20.

Kizilcec, R. F. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *CHI*, 2390–2395.

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 *13th ACM Web Science Conference 2021*, 84–92.

Ledwich, M.; and Zaitsev, A. 2019. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.

Lee, H.-P.; Logas, J.; Yang, S.; Li, Z.; Barbosa, N.; Wang, Y.; and Das, S. 2022. When and Why Do People Want Ad Targeting Explanations? Evidence from a Four-Week, Mixed-Methods Field Study. In *2023 IEEE Symposium on Security and Privacy (SP)*, 923–940.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.

Meta. 2023. How does Facebook decide which ads to show me? . https://www.facebook.com/help/562973647153813. Accessed: 2023-05-20.

Mitchell, T. M.; et al. 2007. *Machine learning*, volume 1. McGraw-hill New York.

OpenCV. 2023. OpenCV Python Library. https://pypi.org/project/opencv-python/. Accessed: 2023-05-20.

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A.; and Meira Jr, W. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 131–141.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *SIGKDD*, 1135–1144.

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*.

Teather, D. 2022. Unofficial TikTok API in Python. https://github.com/davidteather/TikTok-Api. Accessed: 2023-05-20.

TikTok. 2020. How TikTok recommends videos #ForYou. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you. Accessed: 2023-05-20.

TikTok. 2022. Learn why a video is recommended For You. https://newsroom.tiktok.com/en-us/learn-why-a-video-is-recommended-for-you. Accessed: 2023-05-20.

TikTok. 2023. TikTok Terms of Service. https://www.tiktok.com/legal/page/eea/terms-of-service/en#terms-eea. Accessed: 2024-01-10.

Tufekci, Z. 2018. YouTube, the great radicalizer. *The New York Times*, 10(3): 2018.

Wei, M.; Stamos, M.; Veys, S.; Reitinger, N.; Goodman, J.; Herman, M.; Filipczuk, D.; Weinshel, B.; Mazurek, M. L.; and Ur, B. 2020. What twitter knows: characterizing ad targeting practices, user perceptions, and ad explanations through users' own twitter data. In *USENIX Security*, 145–162.

Weller, A. 2017. Challenges for transparency.

Wieringa, M. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *ACM FAccT*, 1–18.

Wilkinson, D.; Namara, M.; Patil, K.; Guo, L.; Manda, A.; and Knijnenburg, B. P. 2021. The Pursuit of Transparency and Control: A Classification of Ad Explanations in Social Media. In *Hawaii International Conference on System Sciences*.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

WSJ Staff. 2021. Inside TikTok's Algorithm: A WSJ Video Investigation. https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477. Accessed: 2023-05-20.

Zannettou, S.; Nemeth, O.-N.; Ayalon, O.; Goetzen, A.; Gummadi, K. P.; Redmiles, E. M.; and Roesner, F. 2023. Leveraging Rights of Data Subjects for Social Media Analysis: Studying TikTok via Data Donations. *arXiv preprint arXiv:2301.04945*.

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, see Section 6.
   (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 4.1.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we listed the data limitations in Section 7. With respect to population-specific distributions, we cannot make any claims about the representativeness of our dataset, given that we perform a sockpuppet-based audit.
   (e) Did you describe the limitations of your work? Yes, see Section 7.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Section 6.
   (g) Did you discuss any potential misuse of your work? No, as we do not anticipate any potential misuse of this work.
   (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 Section 6 and our datasheet in the Supplementary Information.
   (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? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
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)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (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)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
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? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, see the footnote in Section 1.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, as our dataset does not contain personally identifiable information or offensive content.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, see Section 6.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes. We attach the datasheet as a separate document in Supplementary Information.
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? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA