Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Pathways to Conspiracy Theorizing on YouTube in Japan

Koki Ota, Masaki Chujyo, Fujio Toriumi

School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
koki@torilab.net, mchujyo@g.ecc.u-tokyo.ac.jp, tori@sys.t.u-tokyo.ac.jp

Abstract

This study investigates how viewers and creators on Japanese YouTube channels progress towards conspiracy theories. By categorizing channels based on ideological or financial motives and analyzing engagement metrics such as views, likes, and comments, we find that channels driven by monetization, particularly Monetized Conspiracists, promote conspiracy theories more vigorously. This indicates that financial incentives are a crucial factor in the proliferation of such content. Channels that package conspiracy theories in formats like entertainment or spirituality serve as gateways, facilitating viewers' progression towards more extreme conspiracy-laden content. Understanding these pathways is vital for crafting strategies to counteract the spread of conspiracy theories on social media.

Introduction

Conspiracy theories continue to thrive on social media, particularly during the pandemic. Although major platforms have implemented policies to curb misinformation, such as flagging misleading content and removing egregious examples, conspiratorial narratives remain widespread. Even as society has largely returned to near-normalcy with the waning of the pandemic, COVID-19–related conspiracy theories have persisted and evolved, emerging as significant social and political challenges.

Since the outbreak of COVID-19, fake news, misinformation, and conspiracy theories have rapidly proliferated on social media, an issue referred to as an infodemic (Diseases 2020; Zarocostas 2020; Hanley, Kumar, and Durumeric 2023). In particular, analyses of YouTube in 2020 revealed that over a quarter of vaccine related videos contained misleading information, demonstrating that misinformation and conspiracy theories on social media have a significant impact on society (Li et al. 2020). In response to such effects, social media platforms have devised countermeasures, particularly by annotating videos that contain conspiracy theories related to significant social issues, such as vaccines for COVID-19 (YouTube 2023). Additionally, platforms have implemented de-platforming to remove users who spread conspiracy theories. However, it has been shown that while de-platforming can suppress the spread of conspiracy theories, it cannot eradicate them (Innes and Innes 2023).

Understanding why individuals are drawn to conspiracy theories is critical to developing effective interventions. Prior research has identified several factors that contribute to the adoption of conspiracy theories. A number of psychological studies have been conducted, some of which have found a suppressive effect of reflective thinking on conspiracy beliefs about COVID-19 (Yelbuz, Madan, and Alper 2022; Ozono and Sakakibara 2024). In other words, it has been suggested that fostering reflective thinking can be effective in reducing conspiracy beliefs. Moreover, recent studies have highlighted the role of financial incentives on social media platforms, where content creators may be motivated by monetary gains tied to view counts. This economic motivation can drive creators to produce more sensational and conspiratorial content to attract larger audiences and increase ad revenue (Ryan et al. 2020; Ballard et al. 2022).

In this study, we contribute to the prevention of conspiracy theory proliferation by elucidating the pathways through which individuals become engaged in such narratives. Utilizing Japanese YouTube data, we analyze how the prevalence of conspiracy theory content among both content creators and viewers changes over time. While prior research has extensively examined the development of political or ideological radicalization on YouTube (Ribeiro et al. 2020), our study shifts focus to uncover the pathways leading individuals to engage with vaccine-related conspiracy theories. Our main contributions are:

- Our findings indicate that some creators present conspiracy theories as entertainment, associated with more channels than direct conspiracy theory content, suggesting a strategic engagement to boost viewer numbers.
- We identified a subset of creators transitioning from general content to conspiracy theories, showing a noticeable increase in views and likes, highlighting potential financial and engagement incentives for this shift.
- Analysis of viewer transition networks shows that channels combining entertainment or spirituality with conspiracy theory content serve as gateways to more severe conspiracy theories, suggesting these channels play a crucial role in escalating viewer engagement with extreme content.

Category	#Channels	#Videos	#Comments	Avg Videos/Channel	Avg Comments/Channel
Conspiracists	40	2,003	292,639	50.08	7,315.98
Monetized Conspiracists	103	4,164	416,537	40.43	4,044.05
Political Creators	28	825	46,556	29.46	1,662.71
Spiritual Creators	114	4,337	751,993	38.04	6,596.43
Mass Media	71	7,824	2,669,334	110.20	37,596.25
Others	494	12,848	2,271,518	26.01	4,598.22
Total	**850**	**32,001**	**6,448,577**	**37.65**	**7,586.56**

Table 1: Six categories of vaccine-related YouTube channels in Japan, including average videos and comments per channel.

Methodology

Data Collection

To identify vaccine-related channels on YouTube, we used the YouTube Data API to collect data from Japanese-language creators spanning from April 2021 to September 2024. Our search targeted channels whose posted videos included specific vaccine-related japanese keywords commonly associated with conspiracy theories, such as vaccine, ivermectin, Pfizer, Moderna, shedding, and microchip. These keywords were selected to encompass both direct references and slang terms frequently employed in conspiratorial narratives. From the initial set of channels retrieved through these keyword searches, we applied filtering criteria to ensure the relevance and activity of the content creators. Specifically, we retained only those channels with a minimum of 1,000 subscribers and at least 100 total videos. Furthermore, to focus on actively producing channels, we included only those that had uploaded at least 10 videos within the specified data collection period. To maintain the integrity of our dataset and exclude unrelated content, we removed creators who predominantly featured videos unrelated to vaccine conspiracy theories, such as gaming.

Data Annotation

To accurately analyze the role of content creators in disseminating vaccine-related conspiracy theories, we annotated the collected channels based on their relationship to vaccines. The dataset included a diverse range of video types, encompassing vaccine-related news, political commentary, and conspiracy theories. Moreover, we hypothesized that even among creators associated with conspiracy theories, distinct categories exist based on their posting patterns and underlying motivations.

We annotated the channels into the following categories.

1. Conspiracists: These creators are dedicated to propagating conspiracy theories related to vaccines, often presenting unverified claims and misinformation without any apparent intention of monetization. Their content typically lacks credible sources and aims to challenge official narratives.

2. Monetized conspiracists: These creators produce vaccine-related content in an entertainment-oriented manner, such as satirical sketches or sensationalist presentations. Their videos are typically monetized and appear designed to attract wide viewership. A common pattern among these channels is the use of language that avoids explicitly endorsing conspiratorial positions, instead presenting suggestive narratives that may encourage viewers to question mainstream information.

3. Spiritual creators: These creators integrate vaccine-related topics into spiritual or metaphysical content, such as fortune-telling or astrology. They frame vaccines within a mystical context, appealing to audiences interested in spirituality rather than conspiratorial narratives. East Asians, such as Japanese and Chinese, tend to have stronger paranormal and pseudoscientific beliefs (Shiah et al. 2010; Majima et al. 2022), reflected in these creators' content.

4. Political creators: These creators focus on vaccine-related content within a political framework, discussing policies and governmental responses to vaccination programs.

5. Mass media: These channels are created by traditional mass media entities on YouTube, posting TV news broadcasts and similar content.

6. Others: These channels occasionally posted vaccine-related content during the data collection period, but their main focus is on fields adjacent to vaccine-related topics. They typically cover areas such as healthcare, veterinary medicine, or finance, and do not engage with conspiracy theories.

Each channel was manually annotated by a single annotator based on six categories. The annotator reviewed a sample of uploaded videos (titles, thumbnails, and, when necessary, video content), along with video descriptions, channel descriptions, and banner images. Labeling was guided by the predominant themes and presentation styles observed in these materials.

The final annotated dataset comprises 850 channels, distributed across categories as shown in Table 1. Notably, channels associated with conspiracy theories are relatively underrepresented, likely due to the deplatforming of prominent vaccine-related conspiracy theory influencers. Nevertheless, despite platform efforts to curb misinformation, conspiracy theory content still persists, often by circumventing detection and moderation systems.

From Table 1, it is evident that Monetized Conspiracists are more prevalent than Conspiracists in terms of the number of channels (103 vs. 40). However, Conspiracists post more videos and receive more comments on average per channel (approximately 50.08 videos and 7,316 comments) compared to Monetized Conspiracists (approximately 40.43 videos and 4,044 comments per channel). This suggests that while Monetized Conspiracists are more widespread, Conspiracists tend to produce more content and elicit higher engagement per channel. Nonetheless, both groups remain significantly smaller in scale compared to Mass Media channels, which average 110.20 videos and 37,596 comments per channel.

Results

We present two analyses conducted using the vaccine-related YouTube channel dataset: the evolution of conspiracy theory content among content creators and the evolution of conspiracy theory engagement among viewers.

Creators' Pathways to Conspiracy Theorizing

We focus on the categories with the highest degree of conspiracy theory content, specifically Conspiracists and Monetized Conspiracists. We examine the temporal patterns in the conspiracy theory content of videos posted by these channels to understand how their content has evolved over time.

To assess the evolution of conspiracy theory content, we analyzed all videos posted on each channel from its inception until September 30, 2024. Each channel's videos were divided into five equal segments based on the posting order, and within each segment, we calculated the proportion of conspiracy theory videos (hereafter referred to as "conspiracy video rate").

The categorization of each video's conspiracy theory intensity was conducted using a gpt-4o-mini version of ChatGPT (Radford et al. 2018), which classified the content on a 5-point scale based on the video's title and description. This scale was intended to capture the degree of conspiratorial framing, ranging from 1 (clearly non-conspiratorial) to 5 (strongly conspiratorial). Scores of 5 and 4 were categorized as containing conspiracy theories, while scores of 3 and below were considered non-conspiratorial. This threshold was chosen based on an initial qualitative annotation, which indicated that scores of 4 and 5 consistently reflected explicit or strongly suggestive conspiracy narratives.

To validate this classification method, we manually annotated a sample of 100 videos using the same 5-point scale, and subsequently binarized the results using the same threshold (scores of 4 and 5 as conspiratorial, others as non-conspiratorial). This manual evaluation yielded the following accuracy metrics: Precision = 0.90, Recall = 0.98, and F1 Score = 0.94. These results confirm the reliability and sufficiency of the automated classification approach. Due to the large volume of videos, leveraging ChatGPT enabled efficient and consistent measurement of conspiracy theory intensity.

Based on the obtained time series data, we performed time series clustering by the combination of k-means and dynamic time warping (Anh and Thanh 2015), classifying the channels into four distinct patterns of temporal changes in conspiracy video rates. The number of clusters ($k = 4$) was selected using the elbow method, which showed an inflection point at $k = 4$.

Figure 1 illustrates the four clusters of conspiracy video rate trajectories. Each grey line represents an individual channel's trajectory across the five segments, while the red line indicates the cluster center, representing the average pattern within each cluster. Notably, Cluster 2 exhibits a pronounced transition from a low conspiracy video rate of around 20% in the initial segments to over 80% in the final segment. This upward trend suggests that these channels progressively intensify their conspiratorial content, potentially in response to audience engagement strategies and monetization incentives.

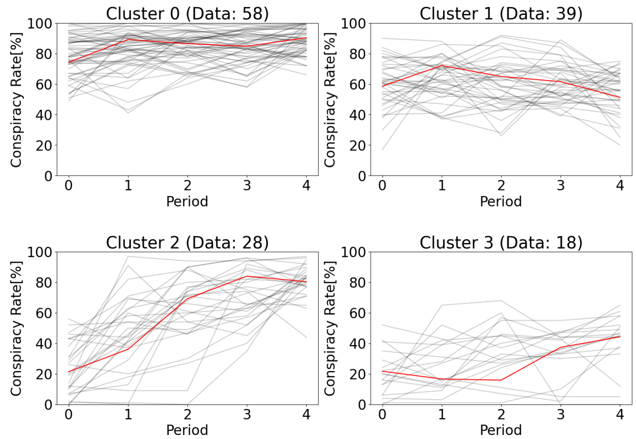

Figure 1: Changes of conspiracy video rates for each cluster.

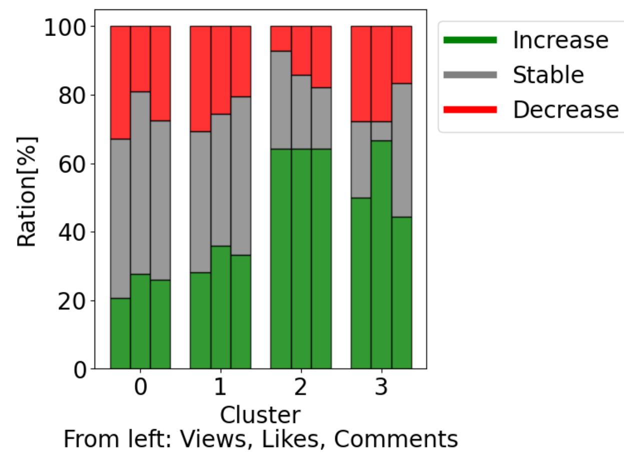

Figure 2: Trend patterns of views, likes, and comments for videos posted in the early and late 40%, from left to right. Each bar represents one of the three metrics, grouped by cluster and time period (early vs. late).

In Figure 2, we illustrate the trends in views, likes, and comments between the early and later periods of each channel. Each group of bars represents these three engagement metrics (from left to right: views, likes, and comments) for the first and last 40% of videos within each cluster. To facilitate comparison, we categorized the changes in each metric into three types based on the difference in median values: Increase, where the median in the latter half is more than twice that of the former half; Decrease, where the median in the latter half is less than half of the former half; and Stable, where neither condition is met. In Cluster 2, the increase trend accounts for 60% of the statistics for each metric, indicating a significant increase in these metrics as the channels transitioned to conspiracy theory content. This trend suggests that shifting to conspiracy theory videos significantly boosts channel activity.

Clusters 0 and 1 consistently exhibit high conspiracy video rates of approximately 80% and 60%, respectively, indicating a persistent focus on conspiratorial content. Conversely, Cluster 3 displays a conspiracy video rate ranging from 20% to 40%, with a modest increasing trend. In Cluster 3, where the conspiracy video rate is ascending, each engagement metric follows an increasing trend, further demonstrating that as the degree of conspiracy theories grows, the views, likes, and comments also tend to rise.

Among the clusters exhibiting a rise in conspiracy content, Clusters 2 and 3 provide a useful contrast: both show increasing conspiracy video rates, but differ in how sharply and substantially they shifted to conspiracy content. This contrast enables us to examine how the magnitude of content shift relates to engagement growth. A t-test on the average change rate of view counts revealed a significant difference between Clusters 2 and 3 ($t = 2.23$, $p = 0.032$), suggesting that the increase in viewership in Cluster 2 may act as a financial incentive that motivates creators to produce more conspiracy content.

Viewers' Pathways to Conspiracy Theorizing

To explore how viewers come to watch conspiracy theory videos, we analyzed their engagements using comment sequences as a proxy for viewing history. To construct a network of channel transitions, we mapped out the sequence of comments users made across different channels. We specifically included edges in the network based on the number of transitions made by viewers from one channel's comment section to another's, requiring a minimum of five transitions between any two channels to ensure a significant connection. Only comments made within a day of each other were considered to capture viewer shifts between channels. From this network, we extracted and analyzed only the largest weakly connected component. The resulting network consists of 300 nodes and 2,518 edges, with an average degree of 16.8, as shown in Figure 3.

To elucidate the pathways by which viewers reach conspiracy theory videos, we implemented random walks on the transition network. Random walks simulate viewer behavior by probabilistically moving between nodes within the network, each representing a channel. Starting from each node, transitions to adjacent nodes were repeatedly executed. These transitions were based on the probabilities derived from the weights of the edges, which represent the number of transitions recorded between each pair of nodes. The random walks were terminated either after reaching the maximum number of steps, set at $L = 10$, or when no further transitions were possible. We conducted 100 random walks from each node.

We calculated the probability of reaching a Conspiracists channel at least once from any of the six categories during these walks. Figure 4 presents the proportion of walks that reached a Conspiracists channel at each step. Among all categories, channels categorized as Conspiracists, Monetized Conspiracists, and Spiritual Creators were notably more likely to be reached to Conspiracists channels. These findings suggest that compared to mass media and other channel types, Monetized Conspiracists and Spiritual Creators act as significant gateways to more severe conspiracy theory content, potentially due to their strategic use of engaging narratives that resonate with broader audiences. Note that Spiritual Creators are largely not included in this network, showing their limited presence.

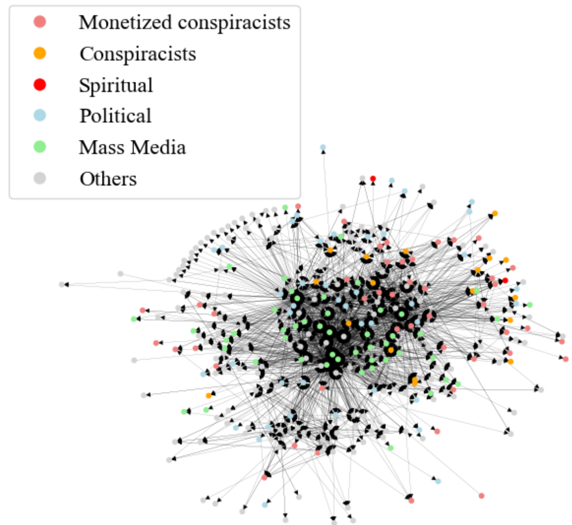

Figure 3: YouTube channel network for viewer transitions.

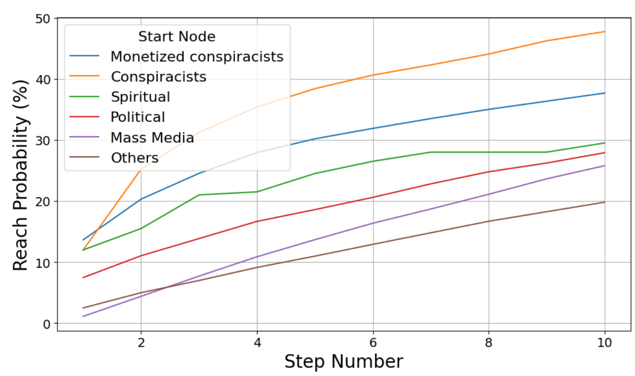

Figure 4: Probability of reaching video of conspiracists.

Discussion and Conclusion

This study collected and manually annotated a dataset of Japanese YouTube channels related to vaccines to examine the extent of their engagement in conspiracy theorizing. We categorized conspiracy theory-related channels based on their motives, ideological or financial, and found that those driven by monetization tend to promote conspiracy theories more vigorously, as evidenced by higher engagement metrics like views, likes, and comments. This suggests that financial incentives are a significant motivator for producing content that can negatively impact society by promoting harmful conspiracy theories.

Our analysis reveals that channels intensifying their conspiracy theory content see increased engagement, indicating that the sensational nature of such content is often leveraged for financial gain. This exacerbates the spread of conspiracy theories. Moreover, channels that present conspiracy theories in more accessible formats, such as entertainment or spirituality, can act as gateways to more severe conspiracy theory content. These channels, particularly Monetized Conspiracists and Spiritual Creators, significantly influence viewers' progression towards more extreme conspiracy theories compared to mass media and other channel types.

Despite these insights, our study faces several limitations. Notably, the more extreme conspiracy theory videos and channels are likely already removed by platform moderation, which can limit the scope of our analysis on social media conspiracy theorizing. Furthermore, we relied on comment transitions as proxies for viewing patterns because more detailed data was not available, which may not fully capture the complex dynamics of viewer behavior. Additionally, the focus on Japanese-language content limits the cultural generalizability of our findings, with behavioral differences potentially influencing how conspiracy theories are presented and perceived in different regions.

Our findings indicate that videos framing conspiracy theories as light entertainment may serve as gateways to more serious conspiracy theory content. However, we do not advocate for the outright avoidance of such themes. Instead, we emphasize the importance of issuing clear warnings to both creators and viewers when addressing themes closely associated with conspiracy theories, to prevent the unintended escalation of their seriousness.

In conclusion, understanding the motivations behind content creation and the dynamics of viewer engagement is crucial for developing targeted strategies to mitigate the spread of conspiracy theories on social media. Our findings suggest that interventions should particularly focus on addressing the incentive structures behind sensational and monetized content, which often serves as an entry point into more harmful conspiracy theorizing.

Acknowledgements

This work was supported by JST RISTEX Japan Grant Number JPMJRS23L1 and JSPS KAKENHI Grant Number JP23H00216.

References

Anh, D. T.; and Thanh, L. H. 2015. An efficient implementation of k-means clustering for time series data with DTW distance. *International Journal of Business Intelligence and Data Mining*, 10(3): 213–232.

Ballard, C.; Goldstein, I.; Mehta, P.; Smothers, G.; Take, K.; Zhong, V.; Greenstadt, R.; Lauinger, T.; and McCoy, D. 2022. Conspiracy Brokers: Understanding the Monetization of YouTube Conspiracy Theories. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 2707–2718. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Diseases, T. L. I. 2020. The COVID-19 infodemic. *The Lancet. Infectious Diseases*, 20(8): 875.

Hanley, H. W. A.; Kumar, D.; and Durumeric, Z. 2023. A Golden Age: Conspiracy Theories' Relationship with Misinformation Outlets, News Media, and the Wider Internet. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Innes, H.; and Innes, M. 2023. De-platforming disinformation: conspiracy theories and their control. *Information, Communication & Society*, 26(6): 1262–1280.

Li, H. O.-Y.; Bailey, A.; Huynh, D.; and Chan, J. 2020. YouTube as a source of information on COVID-19: a pandemic of misinformation? *BMJ global health*, 5(5): e002604.

Majima, Y.; Walker, A. C.; Turpin, M. H.; and Fugelsang, J. A. 2022. Culture as a moderator of epistemically suspect beliefs. *Frontiers in psychology*, 13: 745580.

Ozono, H.; and Sakakibara, R. 2024. The moderating role of reflective thinking on personal factors affecting belief in conspiracy theories. *Applied Cognitive Psychology*, 38(1): e4142.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. *Improving language understanding with unsupervised learning*. Technical report, OpenAI.

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.

Ryan, C. D.; Schaul, A. J.; Butner, R.; and Swarthout, J. T. 2020. Monetizing disinformation in the attention economy: the case of genetically modified organisms (GMOs). *European Management Journal*, 38(1): 7–18.

Shiah, Y.-J.; Wu, M.-H.; Tam, W.-C. C.; and Chang, F. 2010. Paranormal beliefs and religiosity: Chinese version of the Revised Paranormal Belief Scale. *Psychological Reports*, 107(2): 367–382.

Yelbuz, B. E.; Madan, E.; and Alper, S. 2022. Reflective thinking predicts lower conspiracy beliefs: A meta-analysis. *Judgment and Decision Making*, 17(4): 720–744.

YouTube. 2023. Topical context in information panel. https://support.google.com/youtube/answer/9004474?hl=en. Accessed: 2025-01-11.

Zarocostas, J. 2020. How to fight an infodemic. *The lancet*, 395(10225): 676.

Paper 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
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, See Methodology
   (e) Did you describe the limitations of your work? Yes, See Discussion and Conclusion.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, See Discussion and Conclusion.
   (g) Did you discuss any potential misuse of your work? Yes, See Discussion and Conclusion.
   (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
   (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, Our research does not include a hypothesis test.
   (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, Our research does not include a theoretical proof.
   (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, Our research does not include a machine learning experiment.
   (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? Yes
   (b) Did you mention the license of the assets? No, we used the YouTube API and ChatGPT, both of which are open tools. While their use is governed by respective terms of service and licenses, we did not explicitly mention these licenses in the paper.
   (c) Did you include any new assets in the supplemental material or as a URL? No, the study does not release any new datasets or supplementary materials. However, as the data was collected using the YouTube API.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, the study only uses publicly available data obtained through the YouTube API
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, the paper clarifies that the dataset contains publicly available metadata from YouTube and does not include any PII.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? No, the dataset collected in this study will not be made publicly available.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, since the dataset is not intended for public release, a datasheet was not created.
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, Our research does not include crowdsourcing or conducted research with human subjects.

(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

References

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

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.

Ethical Statement

This study uses only publicly available YouTube data and involves no personally identifiable information. The authors declare that they have no competing interests.