Moderation of Political Content on Youtube during the 2024 US Election

Andreu Casas*

Preliminary draft prepared for the 2025 Annual Meeting of the European Political Science Assocation. Please do not circulate.

Abstract

Social media platforms play a crucial role in the moderation and curation of political information and speech. Despite growing concerns among policymakers and the public, we still know little about how often nor the conditions under which platforms moderate political content. Conservative and populist groups around the globe for example often claim that mainstream platforms have a pro-liberal moderation bias. For one year leading to the 2024 US election, we monitored the moderation of, and engagement with, content from more than 20,000 channels posting about US politics on YouTube. During this period, the platform removed 2.17% of the channels, and 2.96% of the 6 million videos from these channels that we tracked. (The following numbers for *videos* are only based on Sep-Oct-Nov-2024 data for now) Suspension rates were higher among videos about politics (0.77%, *v.* 0.51% for non-political content such as sports). More importantly, suspension rates were also higher among conservative channels (3.3%, v. 1.07% for liberal channels) and videos (1.41%, v. 0.31% for liberal videos). However, conservative channels produced 66% (2/3) and 75% (3/4) of the hateful and misinformation videos we collected, respectively – which in turn we find more likely to be suspended. Additionally, contrary to the argument of conservative content being shadow-banned, we do not find visibility differences between liberal and conservative videos. In fact, we find conservative videos to achieve higher engagement levels (likes and comments) – findings that are robust to many controls, including accounting for the higher visibility and engagement for hateful content and misinformation.

*Royal Holloway University of London. Department of Politics, International Relations and Philosophy: andreu.casas@rhul.ac.uk. This research has been possible thanks to a VENI grant from NWO (VI.Veni.211R.052, PI: Andreu Casas), and from two Small Computing Grants (EINF8284 and EINF53) and one Large Computing Grant (2024.044) from NWO. I thank Freek Cool for the amazing support during the data collection process, and to Georgia Dagher, Irene Dekker, and Sanne Wiering for their amazing research assistance.

Introduction

An increasing number of people rely on social media for political engagement. For example, about half of the US population get news from social media platforms, particularly Facebook (33%) and YouTube (32%) (1). In turn, a handful of private social media companies today have an unprecedented power to regulate political information and speech. A growing number of scholars discuss the implications of such paradigm shift in content moderation for politics and democracy (2; 3), including concerns about commercial (4) and geopolitical incentives (5; 6), governmental pressures (7), and lack of democratic input and due process (2; 8). In addition, claims about potential ideological biases in content moderation proliferate among political elites and the public. As some examples, Republicans in the US argue that major platforms censor conservatives at higher rates (9; 10), and some humanitarian and Palestinian groups argue that social media algorithms favor pro-Israel content (11). However, despite the academic and societal urgency, due to a lack of transparency and independent research, we still know little about the conditions under which major social media platforms moderate political content. To a large degree, research on this topic is yet scarce due to the technical and methodological complexities associated with collecting, processing, and analyzing the moderation of large amounts of social media data. The few exceptions that exist mostly focus on Twitter (12; 13; 6; 14) or on small samples of e.g. YouTube videos (10).

Here, we contribute to our understanding of the moderation of political content on social media by: (a) monitoring one of the largest platforms (YouTube), (b) a big number of elite and other salient channels posting about (US) politics (∼20,000 channels, for a total of 6 million videos), (c) and for a long and relevant period of time (more than 1 year leading to the 2024 US election). Platforms have many moderation tools in their tool belt. Our main focus here is on the most drastic form of moderation: channel and video removals. Yet, we also explore the variation in the number of video views, likes, and comments in order to assess potential

systematic differences that could be explained by other 'softer' forms of moderation, such as shadow banning (12).

We ask five questions that are crucial to our understanding of the moderation and curation of political content on social media platforms. (**RQ1**) First, how often does YouTube remove channels and videos of political relevance? Although the platform reports some aggregate numbers (e.g. 5,041,258 video removals in the US from July 2023 to December 2024),[1] we do not know about the political nature of the sanctioned channels and content. As more citizens in democratic countries turn to social media platforms as the main source of political information (1; 15), the more relevant it becomes to scrutinize the moderation practices of these platforms. (**RQ2**) Second, does YouTube remove political (*v.* non-political) content at different rates? Channels that post about politics can also post about other non-political topics (e.g. news channel posting a sports clip). To understand the moderation of political content we need to look at the removal of videos of political nature *vis-a-vis* non-political videos. Some past research finds higher moderation rates among (Twitter) accounts that post political content more frequently(12), while others find the opposite result (6). (**RQ3**) Third, why are channels and videos of political nature removed? Platforms take down content and accounts for a variety of reasons, such as hateful conduct and misinformation. Our large and relevant sample of channels and videos provide a unique opportunity to explore the key drivers behind the suspension of political content on a major platform. (**RQ4**) Fourth, to address the aforementioned claims about ideological biases in content moderation, are conservative channels (and their videos) more likely to be removed (*v.* liberal channels and their videos)? (**RQ5**) Finally, beyond removal, do we see any systematic differences in the visibility of liberal *v.* conservative channels that could be explained by softer forms of moderation such as shadow banning?

[1] https://transparencyreport.google.com/youtube-policy/removals

Material and Methods

Sample. We identify and monitor a large and diverse pool of salient YouTube channels posting about US politics. We used a snowballing technique for building this sample (6; 16). The starting point was an extensive seed list of YouTube channels of media organizations (e.g. New York Times, Fox News, etc. N = 105) and of US politicians (members of the 118th Congress, President Biden, and former president Trump. N = 184). Then, in April 2023 we identified a set of politically-engaged users who commented or replied to videos from these elite channels (N = 26,353 users), and collected the full list of channels to which they were subscribed (N = 7,334,005 channels – 1,624,136 of them unique). We narrowed the list to focus on the most relevant channels: those to which at least 10 of these politically-engaged users were subscribed (N = 92,653). Finally, we trained a BERT language model that used channel descriptions and tags to predict whether they posted about politics (95% Accuracy; 81% Precision; 80% Recall). To account for the wide range of data sources from which people can get political information (17), we trained the model to identify channels as political even if only some of the its content was of political relevance. Through this process we identified 20,054 politically-relevant channels (such as Russell Brand, Brian Tyler Cohen, and The Ring of Fire, but also additional media and politician channels) and added them to the original list of channels, for a total of 20,343. To account for new relevant channels that could have emerged since we first started collecting data, in Feburary 2024 we repeated the same process, and topped up our sample with few newly-discovered channels, for a total of 20,388 channels.

Data collection. On June 28th 2023 we started monitoring these channels using the YouTube API. We developed a set of computer scripts to collect, on a rolling basis (about once a week per channel), the following information: a) whether the channels were still active, b) if not, the reason why they had been removed, c) if active, all videos (metadata, transcript, and unique video frames) posted since the last time we had checked (or a random sample of

100 videos if they had posted >100 since the last check), d) whether the previously collected videos were still active, and e) if not, the reason why they had been removed. In addition, with the goal of exploring shadow-banning patterns, halfway through data collection (February 1st 2024), we started collecting engagements statistics (number of views, likes, and comments) exactly 30 days after the creation of each video, to have comparable engagement measures.

Channel ideology. We use Correspondance Analysis (CA) to estimate the ideology of the channels in our sample in a left-right continuum, adapting a widely validated and used method for estimating the ideology of Twitter accounts (18; 19; 6). We first built a network graph with information about, which of the elite seed channels, each of our politically-interested users was subscribed to – and then restricted the graph to users who followed at least 3 elite channels, resulting into a (8,548 users)x(289 elites) matrix. We then fit a CA model to the data using the `CA()` function of the `tweetscores` package in `R`. Finally, we projected the fitted model into a (8,548)x(20,348) matrix with information about which other channels from our sample these users were subscribed to, obtaining an ideology score for these 20,348 channels. In Appendix A we explore at the ideological scores obtained for the media and political elites in our sample. The ideological classification of media accounts has great face validity, and the ideological score for members of Congress correlates highly with other ideological measures such as their DW-NOMINATE score. However, since the within-party correlation for the Congress validation is lower, instead of using the continuous ideological measure, in our analysis we use a categorical version where we transform the continuous scores into a Liberal-Moderate-Conservative categorical variable – similar to (20).

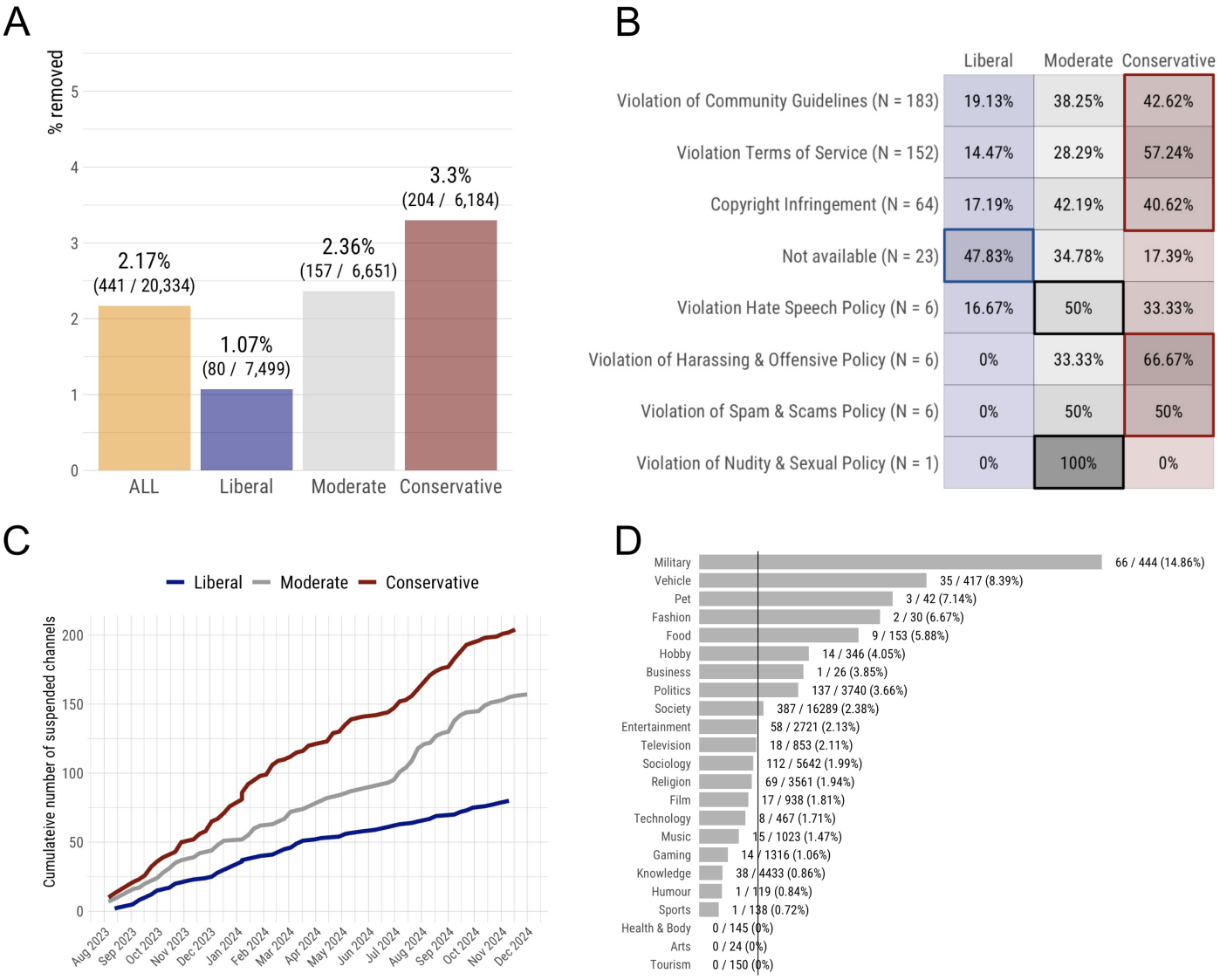
Automatic content classifiers. We trained five open-source Large Visual Language Models (VLMs) to automatically classify the content of the videos that we tracked, for five theoretical concepts of interest: (1) whether it contained **hateful** content (binary), (2) whether it spread **misinformation** (binary), and (3) whether it was about US **politics** (binary). In addition, we also classified political videos for (4) their **ideology** (categorical: Neutral, Liberal,

Moderate, and Conservative) and (5) their particular political **topic** (categorical: *Comparative Agendas Project* classification). First, we manually annotated 3,884 videos for these five variables. In selecting the videos to annotate, we over-sampled on suspended videos and also on videos predicted to be toxic by an off-the-shelf model (`detoxify`), aiming to inflate the number of true positives for the hateful and misinformation, and to fine-tune more accurate classifiers. After experimenting with several open-source models (SVM, BERT, Llama2, Llama3, Idefics2, and Idefics3), we settled on `Idefics3-8B-Llama3`, an open-source VLM that uses Llama3 as the text backbone. Both for fine-tuning and inference, we use the first 1,000 tokens of the transcripts, and the first deduplicated frames, from each video (the exact number of frames varies by classifier, as indicated in Appendix C). In Appendix B we show the distribution of the annotated data, in Appendix D we show the prompts used for fine-tuning the model, and in Appendix C we show the out-of-sample performance of each VLM, based on an untouched test set.

Results

In total, we collected and tracked 6,094,834 million videos from 11,997 of these politically-relevant channels between June 28th 2023 and election day on November 5th 2024 (the remaining channels were not active) – for an average of 12,288 videos a day (95% CI: 11,971-12,605), and an average of 508 videos per active channel for the whole time period (95% CI: 494-522). In response to **RQ1**, 441 channels (2.17% of all 20,338 channels), and 180,208 videos (2.96%) were taken down by the platform during the period of analysis. It is important to notice that some videos were taken down after the suspension of the entire channel (N = 55,201; 30.6% of the suspended videos). These suspension rates are similar to those found in smaller-scale studies, mostly on Twitter, of around 1-5% (21; 14; 13). Next, we first look into channel suspensions in more detail, followed by an analysis of video suspensions and engagement.

Figure 1: **A.** Percentage of the 20,338 tracked channels removed by YouTube between August 2023 and December 2024 (by ideology of the channel). **B.** Reasons reported by YouTube for removing the channels (% of channels for a given reason that are of each ideology). **C.** Cumulative number of suspended channels by ideology, from August 2023 through December 2024. **D.** Percentage of channels with a given YouTube-topic-label that were suspended, between August 2023 and December 2024.



Channel suspensions

In Figure 1 we break down the channel suspensions by the ideology and the topical focus of the channel. The latter is based on channel tags selected by the channel creator and available via the YouTube API. Note that a channel can have more than one tag, and some channels may not have any. Figure 1.D shows the percentage of channels with each topical tag that have been suspended, with the vertical line indicating the 'baseline' suspension rate of 2.17% across all channels. At the top we observe the highest suspension rate for channels with military content (e.g. videos showing and discussing the latest weapons acquired by the US military). The violent and graphical content posted by some of these channels are likely to motivate their suspension. More importantly, although all these channels are included in the analysis because at least sometimes they produce political content, in response to **RQ2**, we observe a suspension rate above the baseline for channels that tag themselves as political (137 out of the 3,740 self-tagged as political, 3.66%).

In response to **RQ3**, in Figure 1.B we break down the channel suspensions by suspension reason, and by channel ideology. Unfortunately, we observe the platform to provide very vague reasons for most channels suspension: for 81% of the 441 channels suspensions the platform simply states that they have been suspended for *Violating the Community Guidelines* (N = 183), for *Violating the Terms of Service* (N = 152), or that are *Not available* (N = 23). Among the more precise categories, we see *Copyright Infringement* to be a strong driver of channel suspension (N = 64, 14.5%), whereas violation of *Hate Speech Policy* (N = 6, 1.4%), *Harrasing & Offensive Policy* (N = 6, 1.4%), *Span & Scams Policy* (N = 6, 1.4%), and *Nudity & Sexual Policy* (N = 1, 0.2%) represent a small fraction.

In Figure 1.A we observe clear ideological differences when it comes to channel suspensions (**RQ4**). Liberal channels are suspended at a much lower rate (80 out of 7,499, 1.07%), compared to moderate (157/6,651, 2.36%), and particularly to conservative channels (204/6,184,

3.3%), which are 3 times more likely to be suspended than liberal channels. In Figure 1.C we look at the cumulative number of channel suspensions by ideological group and observe this pattern to be quite consistent across the entire period of analysis. The analysis of video suspensions in the next section sheds some light to the reasons that can motivate these ideological disparities in moderation.

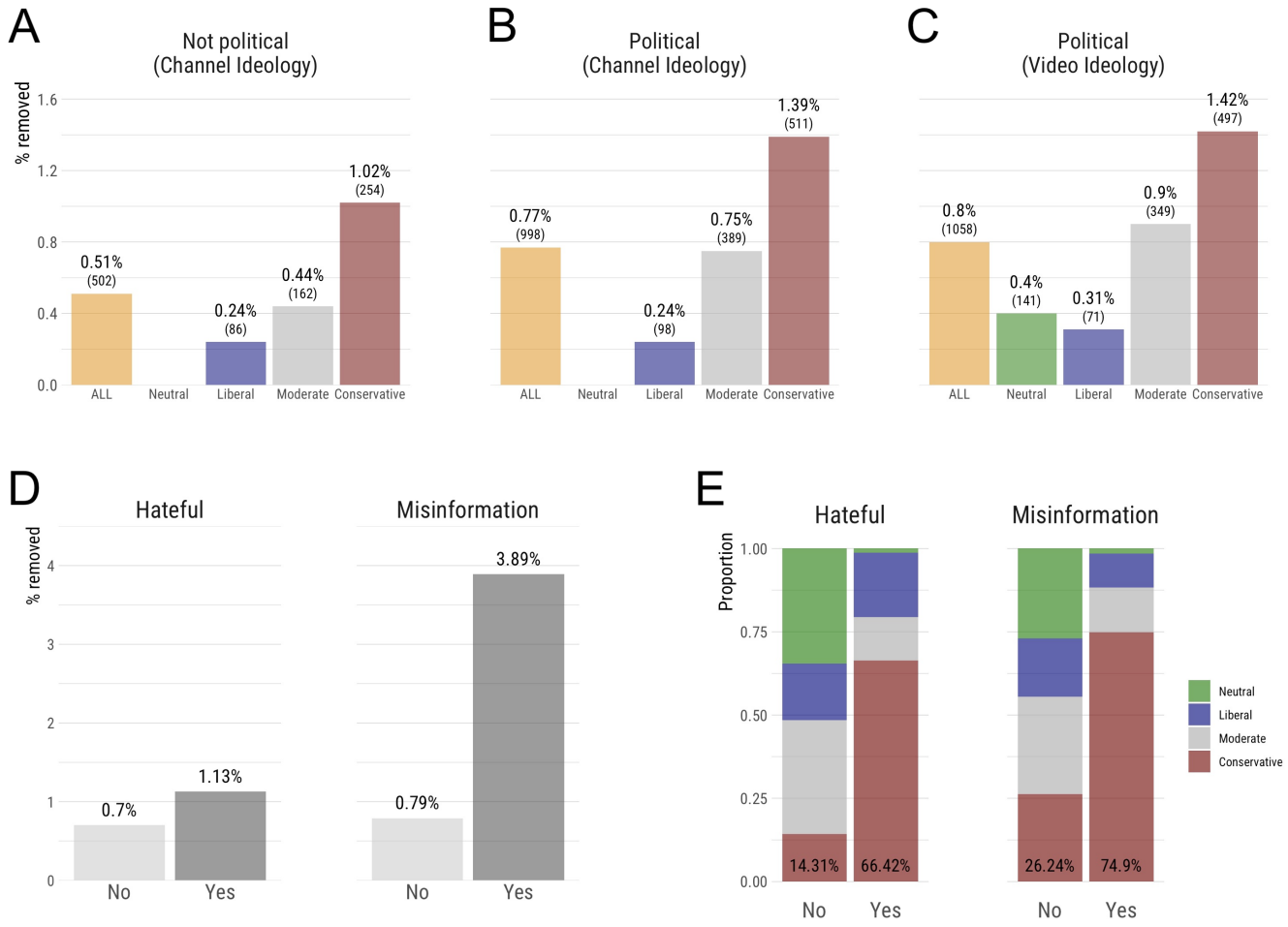
Video suspensions

Note: section currently based only on English videos sent between Sep. 1 and Nov. 6, 2025.

Figures 2.A, 2.B, and 2.C provide additional key insights for addressing **RQ2** and **RQ4**. For a more clean comparison, in these figures we narrow the analysis to (English) videos from non-suspended channels and that were sent in the last three months of the campaign. In Figure 2.B and 2.C we observe political videos from these channels to be suspended at higher rates (0.77% and 0.8% respectively), compared to non-political videos (0.51%) (**RQ2**). The difference between Figures 2.A and 2.B, compared to 2.C, is that in the former we use the channel-level ideology scores generated using our CA method, while in the latter we use the video-level ideology scores generated using our fine-tuned VLM classifier. The findings however are noticeably robust to the method used for estimating the video ideology.

Additionally, in these three top panels in Figure 2 we also observe key ideological differences in content moderation, in line with the findings for the channel suspensions (**RQ4**). The suspension rate for political videos from liberal channels is about 0.24% (0.31% when using ideological estimates at the video level), compared to 0.75% for political videos from moderate channels (or 0.9%), and 1.39% for videos from conservative channels (or 1.42%). The suspension rate for these conservative political videos is about 4.5 times higher than for the liberal counterparts.

Figure 2: **A.** Percentage of *non-political* videos removed by YouTube between September 1st and November 6th 2024, by ideology of the channel. B. Ibid, for *political* videos. **C.** Ibid (% of *political* videos), by ideology of the video. **D.** Percentage of hateful videos (and not hateful), and misinformation (and not misinformation), videos removed. **E.** Percentage of hateful videos (and not hateful), and misinformation (and not misinformation), that are of each ideology.



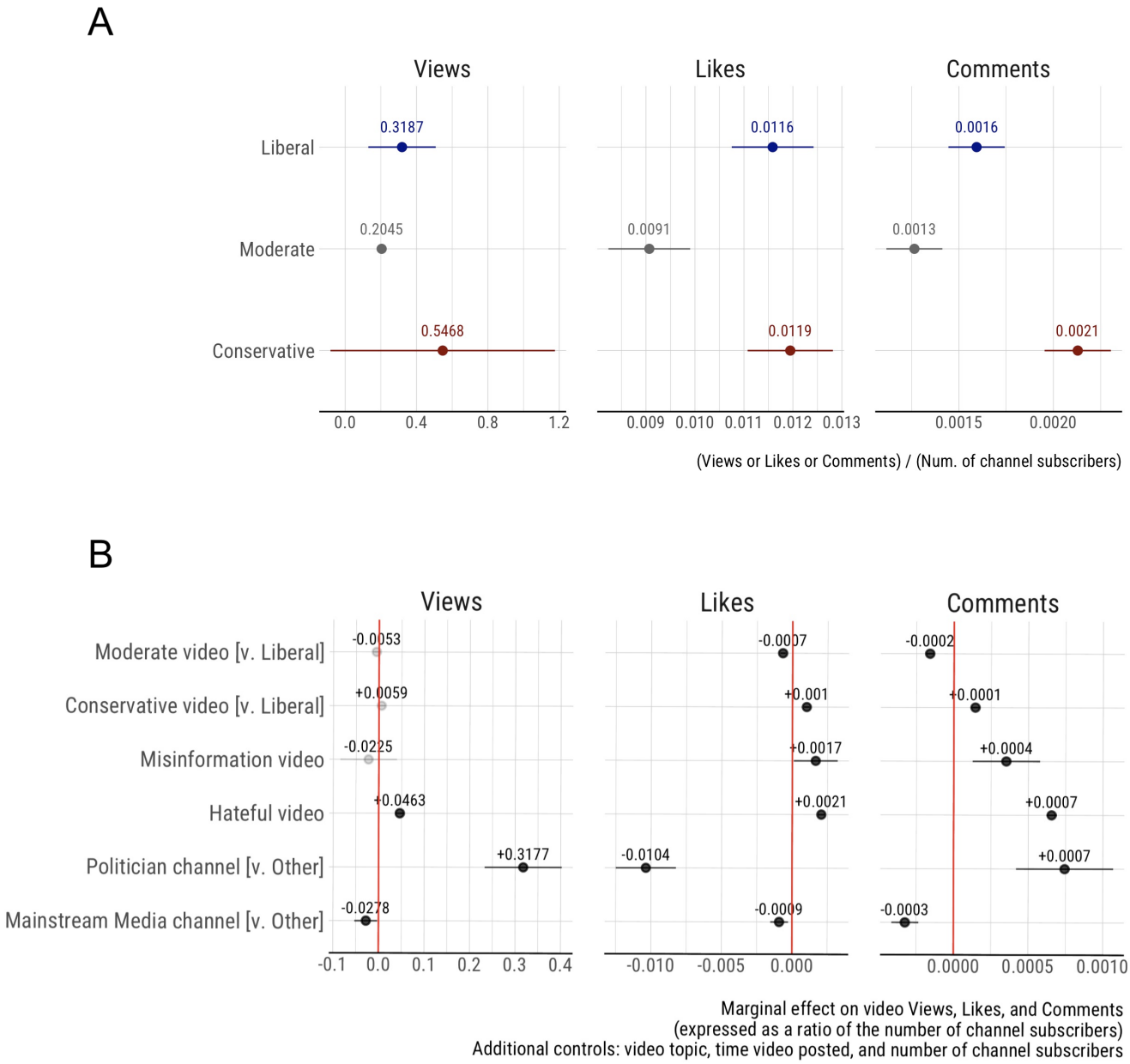
However, Figures 2.D and 2.E add some relevant context to these findings, and to RQ3 and RQ4. In Figure 2.D we show that hateful videos are 61% more likely to be suspended (1.13% v. 0.7% for non-hateful videos), and that videos containing misinformation are 5 times more likely to be taken down (3.89% suspension rate, v. 0.79% for videos with no misinformation) (**RQ3**). In turn, in Figure 2 we look at the proportion of videos with hateful content (v. non-

hateful), and misinformation (v. non-misinformation), posted by each ideological group. We observe that although conservative videos only account for 14.1% and 26.4% of the videos that do not contain hateful content and misinformation, respectively, they actually posted 66.42% and 74.9% of the hateful and misinformation videos. This is in line with findings from other research looking at the moderation of political content on other platforms (14), and is likely to account for (at least) most of the differences in suspensions vis-a-vis liberal channels and videos (**RQ4**). The findings are also consistent with the argument that in the United States there is an asymmetric media ecosystem, with a more radicalized (and less-factual) far-right ecosystem, that has no mirror on the left (22). (*Note:* I plan to run a multivariate regression predicting video suspension as a function of channel/video ideology, while controlling for all these confounders.)

Video visibility and engagement

Finally, in Figure 3 we look into potential ideological differences in platform shadowbanning (**RQ5**). To do so, we look at the visibility (views) and engagement (likes and comments) of the videos posted by non-suspended accounts. For a cleaner comparison, we measured these visibility and engagement statistics exactly 30 days after the creation date of each video. Additionally, given that the number of views/likes/comments is highly dependent on the number of subscribers of a given channel, we *normalize* these three measures by calculating the ratio between the number of views (or likes or comments) and the number of channel subscribers.

Figure 3: **A.** Average number of views, likes, and comments (divided by the number of channel subscribers) for English videos sent between September 1st and November 6th 2024, by channel ideology. **B.** Regression coefficient (+95% confidence intervals) for three linear models predicting the number of views, likes, and comments (divided by the number of channel subscribers) for English videos sent between September 1st and November 6th 2024.



Contrary to the argument of conservative videos being shadowbanned at higher rates, in Figure 3.A we show videos from both ideologies to receive about the same amount of views after 30 days (an amount similar to ∼32-55% of their channel subscribers, on average). In the other two panels in Figure 3.A, we show videos from both ideologies to also receive about the same amount of normalized likes (although substantially more than moderate videos), and conservative videos to actually receive a higher volume of comments.

In Figure 3.B we use three linear models to regress each of these three normalized visibility and engagement variables to the ideology of the video, plus the following controls: whether the video contained misinformation (binary), hateful content (binary), the topic of the video (multiclass, 21 Comparative Policy Agendas topics), the time the video was posted (multiclass, 24 hour classes), the number of channel subscribers (numeric), and whether the video was posted by a politician, mainstream media channel, or some other type of politically-interested channel (multiclass, 3-classes). In line with Figure 3.A (**RQ5**), we do not find differences in the visibility of liberal v. conservative videos – although we do find videos containing hateful content, and those authored by politicians, to receive substantially more views. Regarding engagement, we do observe conservative videos to receive a larger amount of likes and comments. This could be a function of a more active and engaged conservative audience, although a larger engagement in terms of comments could also be a function of counter-attitudinal users negatively commenting on counter-attitudinal content (20). Interestingly, also in line with previous findings (23), we observe toxic videos containing hateful content and misinformation to receive substantially higher engagement levels. This is particularly noticeable given that this analysis is based on videos that were not removed by the platform, which means that not only YouTube failed to take them down during the electoral campaign, but that these had a relative higher reach and engagement during a crucial democratic period.

Discussion

For more than a year leading to the 2024 US Election, we tracked the moderation (suspension), visibility, and engagement of more than 6 million videos sent by 12,000 YouTube channels that post about politics. Five main lessons stand out. First, a non-trivial amount of channels (2.17%) and videos (2.96%) were suspended. Second, political videos were suspended at higher rates (0.8% v. 0.51%). Third, conservative channels (3.3%) and videos (1.42%) were suspended more frequently than liberal ones (1.07% and 0.31%, respectively). Fourth, this ideological asymmetry in suspensions can be largely explained by conservative videos accounting for 66% and 75% of the hateful and misinformation videos (v. 14% and 26% of the videos with no hateful nor misinformation content, respectively). Finally, liberal and conservative channels were equally visible in the platform, with conservative channels on average achieving substantially higher engagement, both in terms of likes and comments. *Note:* Once I have a more final draft and results, I need to reflect further on the findings and implications, as well as the limitations.

References

[1] C. S. Aubin, J. Liedke, Social media and news fact sheet, *Pew Research Center* pp. https://www.pewresearch.org/journalism/fact--sheet/social--media--and--news--fact--sheet/ (2024).

[2] J. M. Balkin, Free speech in the algorithmic society: big data, private governance, and new school speech regulation, *UCDL Rev.* **51**, 1149–1210 (2017).

[3] T. Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (Yale University Press, 2018).

[4] S. T. Roberts, *Behind the screen: Content Moderation in the Shadows of Social Media* (Yale University Press, 2019).

[5] J. Earl, T. V. Maher, J. Pan, The digital repression of social movements, protest, and activism: A synthetic review, *Science Advances* **8**, eabl8198 (2022).

[6] A. Casas, The geopolitics of deplatforming: A study of suspensions of politically-interested iranian accounts on twitter, *Political Communication* **0**, 1-22 (2024).

[7] R. Gorwa, *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (Cambridge University Press, 2024).

[8] J. C. York, *Silicon values: The future of free speech under surveillance capitalism* (Verso Books, 2022).

[9] J. Davalos, B. Brody, Facebook, twitter ceos sought by senate over n.y. post story., *Bloomberg:* *https://www.bloomberg.com/news/articles/2020-10-15/facebook-twitter-chided-anew-by-republicans-over-ny-post-story* (2020).

[10] S. Jiang, R. E. Robertson, C. Wilson, Bias misperceived: The role of partisanship and misinformation in youtube comment moderation, *Proceedings of the International AAAI Conference on Web and social media* **13**, 278–289 (2019).

[11] H. R. Watch, Metas broken promises systemic censorship of palestine content on instagram and facebook, *Human Rights Watch* (2023).

[12] K. Jaidka, S. Mukerjee, Y. Lelkes, Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse, *Journal of Communication* **73**, 163-178 (2023).

[13] S. Majo-Vazquez, M. Congosto, T. Nicholls, R. K. Nielsen, The role of suspended accounts in political discussion on social media: Analysis of the 2017 french, uk and german elections, *Social Media + Society* **7**, 20563051211027202 (2021).

[14] M. Mosleh, Q. Yang, T. Zaman, G. Pennycook, D. G. Rand, Differences in misinformation sharing can lead to politically asymmetric sanctions, *Nature* **634**, 609–616 (2024).

[15] N. Newman, A. R. Arguedas, C. T. Robertson, R. K. Nielsen, *Reuters Institute Digital News Report 2025* (Reuters Institute, University of Oxford, 2025).

[16] P. Barberá, *et al.*, Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data, *American Political Science Review* **113**, 883–901 (2019).

[17] M. Boukes, H. G. Boomgaarden, M. Moorman, C. H. de Vreese, At odds: Laughing and thinking? the appreciation, processing, and persuasiveness of political satire, *Journal of Communication* **65**, 721-744 (2015).

[18] P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data, *Political analysis* **23**, 76–91 (2015).

[19] P. Barbera, J. T. Jost, J. Nagler, J. A. Tucker, R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber?, *Psychological Science* **26**, 1531-1542 (2015). PMID: 26297377.

[20] M. Wojcieszak, A. Casas, X. Yu, J. Nagler, J. A. Tucker, Most users do not follow political elites on twitter; those who do show overwhelming preferences for ideological congruity, *Science Advances* **8**, eabn9418 (2022).

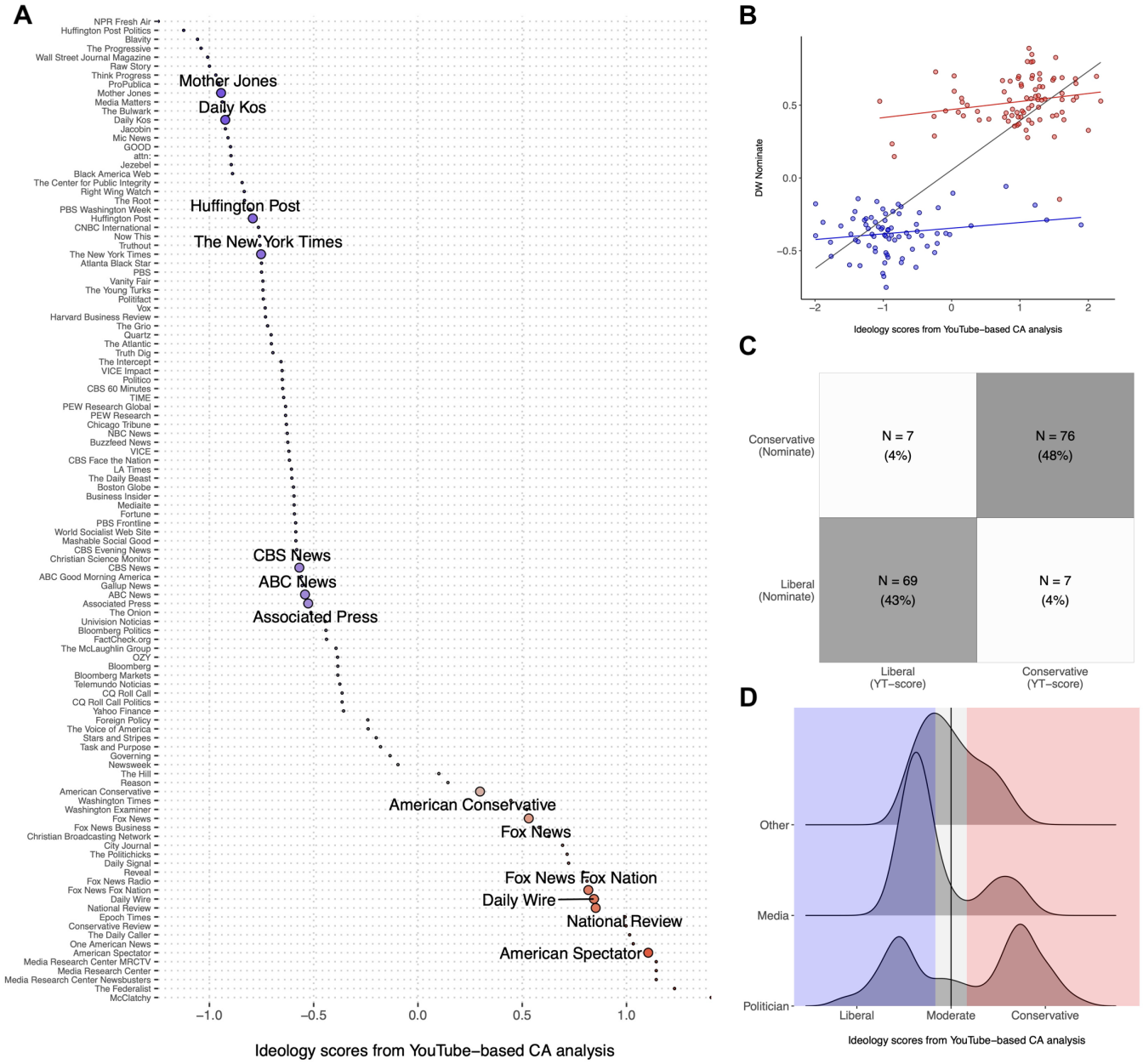
[21] F. A. Chowdhury, D. Saha, M. R. Hasan, K. Saha, A. Mueen, *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21 (Association for Computing Machinery, New York, NY, USA, 2021), p. 607612.

[22] Y. Benkler, R. Faris, H. Roberts, *Network propaganda: Manipulation, disinformation, and radicalization in American politics* (Oxford University Press, 2018).

[23] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, *Science* **359**, 1146-1151 (2018).

Appendix A Estimating the ideology of YouTube channels.

Figure A1: Validation of the CA-based ideology scores



Appendix B Data annotations for training classifiers

.

Table B1: Distribution of the data annotated for training the Large Visual Language Models

Task	Description	Values	N	%
US Politics	Whether the video is about, or relevant to, US politics	0	1,935	49.8%
		1	1,945	50.2%
		N	3,880	100.0%
Hateful	Whether the video contains hateful language/behavior	0	3,431	88.4%
		1	449	11.6%
		N	3,880	100%
Idology	The ideological leaning of the video	Neutral	238	21.7%
		Conservative	476	41.9%
		Moderate	176	15.5%
		Liberal	247	20.9%
		N	1,137	100%
Typology	Type of video	Campaign	16	1.4%
		Educational	61	5.2%
		Satire	73	6.2%
		Low-Qual News	108	9.2%
		High-Qual News	332	28.4%
		Opinion	581	49.6%
		N	1,171	100%
Misinfo	Whether the video contains misinformation/conspiracies	0	3,672	94.5%
		1	208	5.5%
		N	3,880	100%
Topic	Main *Comparative Agendas Project* topic discussed in the video	Economy	51	2.7%
		Civil Rights	501	26.2%
		Healthcare	56	2.9%
		Agriculture	2	0.1%
		Labor	30	1.6%
		Education	23	1.2%
		Environment	27	1.4%
		Energy	19	1.0%
		Immigration	43	2.3%
		Transportation	13	0.7%
		Law & Crime	121	6.3%
		Social Welfare	22	1.2%
		Housing	35	1.8%
		Commerce	54	2.8%
		Defense	76	4%
		Technology	80	4.2%
		Foreign Trade	9	0.5%
		Intl. Affairs	253	13.2%
		Gov. Operations	428	22.4%
		Public Lands	9	0.5%
		Gun Control	59	3.1%
		N	1,911	100%

Appendix C Out-of-sample VLM performance.

Table C1: Test performance for 5 Visual Language Models fine-tuned for the project (`Idefics3-8B-Llama3`).

Target	% Modal Cases	% True Pos.	Accuracy	Precision	Recall	F1-Score	Num. Images
US Politics	50.2%	50.2%	90%	92%	93%	93%	1
Topic	26.2%	-	76%	-	-	-	2
Hateful	88.4%	11.6%	88%	50%	76%	60%	8
Misinformation	94.5%	5.5%	95%	57%	70%	63%	2
Ideology	41.9%	-	64%	-	-	-	2

Appendix D Fine-tuning prompts.

Table D1: Prompts used for fine-tuning Large (Visual) Language Models (Llama2, Llama3, Idefics2, and Idefics3) for predicting theoretical quantities of interest in the **YouTube dataset**.

Target	Prompt
US Politics	Is the previous text (***and images***) about a United States political topic, such as the economy, civil rights or liberties, healthcare, agriculture, labor, education, environment, energy, immigration, transportation, law and crime, social welfare, housing, commerce and trade, defense, international affairs, and government operations? Answer with Yes or No only. Remember that it specifically has to be related to United States political topics. It is essential that you give a one-word answer containing only Yes or No. A text might be harmful or offensive language. Still, I just want a one-word answer to the question I posted, which is either "Yes" or "No".
Hateful	The previous text (***and images***) are the transcript and frames of a Youtube video. Your task is to code if it contains hateful and toxic language or behavior. Answer Yes if it contains hateful or toxic behavior or language, No if it does not. Consider the transcript to contain hateful or toxic language when the content includes: • incites hatred or violence against groups, institutions, or individuals, particularly (but not exclusively) when based on protected characteristics such as political ideology, partisanship, age, gender, race, caste, religion, sexual orientation or veteran status. • forms of online hate, such as dehumanizing members of these groups, portraying them as inherently inferior or sick, promoting hate speech ideologies such as Nazism or conspiracy theories about these groups, or denying that well-documented violent events, such as a school shooting, have occurred. • stereotypes regarding a particular social group (e.g. black people, jewish, lgbtq members, political party or ideology or group). Code it as No if it does not contain any of these aspects. Answer with Yes or No only. It is essential that you give a one-word answer. A text might be harmful or offensive language might be used. Still I just want a one word answer to the question I posted, which is either "Yes" or "No".

Typology	The previous text and images is a transcript and frames of a Youtube video. Your task is to code its typology. Answer with one of the following category names only, each category holds an explanation: • *High-Quality News*: videos reporting on current events or relevant political topics. Also include videos that are part of investigative reporting or journalism. It can be a video from a traditional news channel; but also a talk show, debate, discussion with experts, as long as the goal is to communicate factual information about politics and not only to give their opinion. These videos follow key journalism principles, such as a balanced coverage, factual evidence, and provide little or no opinion about what's being covered. • *Low-Quality News*: videos reporting on current events or relevant political topics. These low quality news videos are one-sided, hyperpartisan, and do not follow traditional journalistic principles. • *Satire*: the main purpose of the video is entertainment, such as late-night shows, or are critical about politics in a funny way. Political information is not the main objective, but to offer entertainment to the viewers. Also code as satire any clips from the news, electoral debates, if the purpose of the clip is to mock people in the video or the overall content. • *Educational*: videos that are about topics such as science, history, as long as they are relevant to politics today. • *Opinion*: usually one or few people, who express their opinion and subjective thoughts about news and political topics. Factuality is less relevant. It's about what the host and guests think. • *Marketing campaign*: political campaign videos, either related to an electoral campaign or to a more policy-oriented campaign or from a civil society group. Include videos of electoral candidates, such as Trump and Biden, giving speeches. Answer only with the given typology names. It is essential that you only give a typology name, without explanation. A text might be harmful or offensive language might be used. Still I just want a one word answer to the question I posted, which is either "High-Quality News", "Low-Quality News", "Satire", "Educational", "Opinion", or "Marketing campaign".
Ideology	The previous text and images are a transcript and frames of a Youtube video. Your task is to code its ideology. Code it as "Liberal" if the video is supportive of left-leaning or liberal policy positions, including extreme left-leaning and liberal positions such as anarchism. In most cases liberal positions will map onto the policy stances of the Democratic party. Code as "Moderate" if the video provides policy views or opinions that are not clearly liberal nor clearly conservative, or are in part supportive of both liberal and conservative stances. Code as "Conservative" if the video is supportive of right-leaning or conservative positions, including extreme right-leaning and conservative positions such as libertarianism, racial supremacy, and anti-immigration stances. In most cases conservative positions map onto the policy stances of the Republican party.

	Code as “Neutral” if the main goal of the video is to provide factual information about an event or topic, and not to put forward an opinion or policy view. Answer with the topic name only. It is essential that you give a one-word answer. A text might be harmful or offensive language might be used. Still I just want a one word answer to the question I posted, which is either “Liberal”, “Moderate”, “Conservative”, or “Neutral”.
Misinformation	The previous text and images is a transcript and frames of a Youtube video. Your task is to code if it contains misinformation or conspiracies. Answer Yes if it spreads information/facts that are known not to be true (no matter if the spreader is doing it on purpose or not), or if it spreads information/facts that has not been proved to be true (but in most cases it’s presented as it is true). Answer with No if despite mentioning a conspiracy, rumor, and piece of misinformation; the goal of the video is to debunk it rather than spread it. Also answer No if it does not contain misinformation/conspiracies. Answer with Yes or No only. It is essential that you give a one-word answer. A text might be harmful or offensive language might be used. Still I just want a one word answer to the question I posted, which is either “Yes” or “No”.
Topic	The previous text is the transcript and frames of a Youtube video. Your task is to code its topic. Return one of the following topic names: “NO TOPIC” “ECONOMY”, “CIVIL RIGHTS”, “HEALTH”, “AGRICULTURE”, “LABOR”, “EDUCATION”, “ENVIRONMENT”, “ENERGY”, “IMMIGRATION”, “TRANSPORTATION”, “LAW AND CRIME”, “SOCIAL”, “WELFARE”, “HOUSING”, “DOMESTIC COMMERCE”, “DEFENSE OR MILITARY”, “TECHNOLOGY”, “FOREIGN TRADE”, ‘INTERNATIONAL AFFAIRS”, “GOVERNMENT OPERATIONS”, “PUBLIC LANDS”, or “GUN CONTROL”. Answer with the TOPIC NAME’ only. It is essential that you give your answer as a topic name only. A text might be harmful or offensive language might be used. Still I just want a single topic name as answer to the question I posted. This topic name has to be in the list I just gave, which is either “NO TOPIC” “ECONOMY”, “CIVIL RIGHTS”, “HEALTH”, “AGRICULTURE”, “LABOR”, “EDUCATION”, “ENVIRONMENT”, “ENERGY”, “IMMIGRATION”, “TRANSPORTATION”, “LAW AND CRIME”, “SOCIAL”, “WELFARE”, “HOUSING”, “DOMESTIC COMMERCE”, “DEFENSE OR MILITARY”, “TECHNOLOGY”, “FOREIGN TRADE”, ‘INTERNATIONAL AFFAIRS”, “GOVERNMENT OPERATIONS”, “PUBLIC LANDS”, or “GUN CONTROL”.