YouTube's Adpocalypse - Study of creators' participation on the platform after the implementation of a demonetisation policy

Maria Rasskazova
ACT and inIdEx ICCA
University Sorbonne Paris Nord
m.rasska@gmail.com

Job Market Paper[*].

This version: January 2026.

Abstract:

Platforms may implement diverse incentives and rules to attract users and shape their interactions. These tools can be adjusted over time in response to the massive departure of users. This paper focuses on the events that occurred on YouTube at the beginning of 2017. The massive departure of advertisers from YouTube in March 2017 led to a shift in the platform's moderation policy around video monetisation. This work examines the causal effect of the implementation of a broader demonetisation rule on content supply of 51,583 English-speaking channels on YouTube.

This research offers insights into the impact of an economic sanction on creators' production strategies on two levels: the number of weekly posted videos and the proportion of not-suitable-for-advertisers content in the weekly video supply. First, the adjustment in the demonetisation rule led to a decrease in the volume of weekly video supply by non-advertiser-friendly content creators compared to brand-safe ones. Second, the impact of such moderation action is heterogeneous among the channel's main content categories and its audience size. Third, the introduction of this new rule led to a reduction of the share of non-advertiser-friendly videos in the volume of weekly supply.

These findings provide additional evidence on the importance of extrinsic motivations and financial incentives in the creators' level of participation on the platform. These results illustrate the platform's ability to control the flux of content. Moreover, it highlights the limits of a uniform moderation action on the reduction of harmful content and as a platform governance strategy.

Keywords: Platform governance, content moderation, advertisers, demonetisation, content creators, YouTube.

---

[*] *Acknowledgements:* This research is supported with funding from the inIdEx Industries Culturelles & Creations Artistiques (inIdEx ICCA). I am grateful to Ambre Elsas-Nicolle for her continuous advice and support. This paper greatly benefited from discussions with and comments made during the 2024 AFREN Summer School.

Associated Jupyter notebooks and R scripts to this research are available on GitHub: https://github.com/mrasska/YT_Demonetization_policy_analysis

Introduction

On December 1st 2017, Felix Kjellberg, also known as PewDiePie, released a video entitled *"YouTube's secret uncovered!"* in which he discussed the main changes in YouTube's content monetisation policy. As PewDiePie, many creators expressed their concerns over the modifications made in YouTube's guidelines in 2017 following advertisers' massive departures. In response, the platform announced several changes in its guidelines [2] to create a more advertiser-friendly (brand-safe) environment. Brand-safe rules aim to shape content and interactions to brands' ideal, i.e., as non-controversial, non-political, non-sexual and non-vulgar (Fahey 1991 as cited in Bishop 2021).

These implemented changes are part of platform governance mechanisms. In order to retain users on one side and/or encourage specific behaviour, a platform may change its design and/or rules and place itself as "*a focal, private regulator*" (Boudreau & Hagiu, 2009). Among the variety of available instruments, moderation is a non-pricing technique "*that structures participation in a community to facilitate cooperation and prevents abuse*" (Grimmelmann, 2015). Diverse actions are available to regulate interactions among users and to control published content (Jiang et al., 2023). Some of them may sanction users whose behaviour violates the rules. Besides deleting posts and withdrawing users' access, i.e., removing the rule-violating action, platforms can adjust the rules around content monetisation, by creating economic sanctions.

Financial rewards and sanctions are another type of moderation measure used by content-sharing platforms, as they may influence the participation and behaviour of content creators. Content producers may pursue their activity for intrinsic motivations, such as self-expression, sharing knowledge and pursuing hobbies (Zimmer & Scheibe, 2019). However, extrinsic motivations, like building a reputation or earning income, may justify the time investment in this activity (Törhönen et al., 2019). Previous work highlighted the positive influence of introducing economic incentives on creators' content supply, as it led to an increase in the volume and quality of the publications (Sun & Zhu, 2013; Tang et al., 2012). However, very little is currently known about the effects of introducing economic sanctions on users' activity.

The changes introduced by YouTube following advertisers' departure in March 2017 enable researchers to assess the influence of financial sanctions on content producers' participation, more precisely on the evolution of the volume of their publications. On YouTube, content monetisation

[2] Bardin A (2017) Strengthening YouTube for advertisers and creators. In: YouTube Official Blog. https://blog.youtube/news-and-events/strengthening-youtube-for-advertisers/.

Schindler P (2017) Expanded safeguards for advertisers. In: Google - The Keyword. https://blog.google/technology/ads/expanded-safeguards-for-advertisers/.

relies on two non-excludable eligibility criteria: meeting the platform-partnership requirements and following advertiser-friendly content guidelines. As a result of advertisers' departure from the platform, the latter changed both eligibility requirements. While some researchers have investigated the causal effect of the adjustments to partnership criteria on creators' participation (Abou El-Komboz et al., 2023; Andres et al., 2023), the impact of the introduction of a demonetisation policy on the evolution of user-generated content provision is less clear.

Video demonetisation consists of removing the ability to earn advertising revenue while the published content remains available on the platform (Griffin, 2023). The criteria for this sanction are publicly available in a platform's "advertiser-friendly content" guidelines. Dunna et al. (2022) examined the influence of this measure on creators' activity. They found that this type of sanction leads to a decrease in the number of views for the targeted content. The loss of visibility and revenues may demotivate some creators. Moreover, as creators' economic models, i.e. their production and monetisation strategies, are diverse, the causal effect of introducing a financial sanction may differ from one content producer to another (Feher, 2023). This paper aims to address the issue: ***how does the enforcement of a demonetisation policy affect creators' participation on a social media platform?***

To answer this research question, we examine the weekly content supply of 51,583 English-speaking channels on YouTube. Using a difference-in-differences approach, we estimate the causal effect of the implementation of a broader demonetisation sanction on creators' activity. To do so, we draw a distinction between channels that post content around themes that may violate the advertiser-friendly content guidelines (treatment group, also referred to as non-advertiser-friendly channels) and those that comply with advertisers' standards (control group, also referred to as brand-safe producers). As there were multiple community guideline adjustments in 2017, we investigate the impact of YouTube's first statement published in March 2017 that introduced a "*broader demonetisation*" policy [3] and we analysed users' participation between August 2016 and October 2017.

This research is closely related to Abou El-Komboz et al. (2023) and Andres et al. (2023). In their work, they analysed the causal effect of the new YouTube Partnership Programme eligibility criteria, introduced after the 2017 consecutive advertisers' massive departures on creators' activity. Abou El-Komboz et al. (2023) found a decrease in creators' activity (frequency, volume and quality

[3] Bardin, A. (2017, March 20). Strengthening YouTube for advertisers and creators. *blog.youtube*. https://blog.youtube/news-and-events/strengthening-youtube-for-advertisers/

of video uploads) on YouTube after the platform's intervention. Concurrently, Andres et al. (2023) found an increase in creators' publication volume on competing platforms.

This present study sheds additional light on financial incentives in content supply strategies. The adjustments in the eligibility criteria can be defined as an entry barrier in creators' monetisation strategy (Kopf, 2020), whereas demonetisation concerns a limited set of videos. This research provides empirical insights into this specific economic penalty. We find that the introduction of the demonetisation policy leads to a statistically significant decrease in content supply from non-advertiser-friendly creators compared to brand-safe ones. It also affects the share of unsuitable-for-advertisers content in the weekly content supply of exposed producers. These findings confirm the influence of financial incentives on users' engagement on the platform (Kerkhof, 2024; Törhönen et al., 2019). Economic sanctions as part of a moderation strategy influence the behaviour of users and shape the content available on the platform (Zeng & Kaye, 2022).

However, the effectiveness of a uniform action may vary from one creator to another. Treatment effect heterogeneity is observed across the channel's main content category and its audience size (number of subscribers). This highlights the limits of uniform moderation measure. Moreover, this reveals the diversity of creators' strategies, more precisely, how content creators adapt their production in response to new platform policies. Previous research tended to study producers' side as a whole, without underlining the differences among creators. As producers' characteristics differ across individuals, this research provides additional insights into the complex dynamics in content creation and moderation between creators and platforms.

Finally, this work highlights the implications of ad-funded business models of media platforms on cultural production. By being dependent on advertising revenues, platforms implement advertisers-driven moderation policies (Choi & Jeon, 2023). The latter may not be without consequences on creators' participation. This paper points out the influence of such policies on cultural production, specifically on the volume of published user-generated content.

This paper is structured as follows. Section 1 gives a brief overview of YouTube's content moderation and monetisation policies. The second section presents a summary of the literature on content moderation and its influence on users' participation. Section 3 describes the methodology used for this study. Section 4 presents the econometric model used to estimate the causal effect of the demonetisation policy on creators' content supply. Section 5 presents the results of the estimations. Section 6 explores the mechanisms behind creators' production strategies. Section 7 concludes.

Section 1: Industry background

1.1. YouTube monetisation and moderation policy

Monetisation refers to the conversion of user traffic into revenues (Goanta et al., 2022). On YouTube, content creators can earn revenue from the platform's internal monetisation system. However, this monetisation option is subject to two eligibility criteria: being a member of YouTube Partnership Programme (YPP) and following YouTube's community guidelines.

In 2012, YouTube introduced its partnership programme as an incentive for its creators to produce high-quality content [4]. This programme allows its members to monetise their videos by enabling advertisement placements on and around published content. The revenues from the advertisements are shared between YouTube and partnered creators. The initial version of the partnership programme was open to all users, regardless of the number of views or subscribers [5]. After the multiple massive departures of advertisers from the platform, various eligibility changes have been implemented, such as metric thresholds.

In addition to the YPP, the published video needs to be advertiser-friendly. YouTube defines a non-advertiser-friendly video, i.e., content that is not eligible for monetisation, as follows:

> *"Content that is considered "not advertiser-friendly" includes, but is not limited to:*
> *•Sexually suggestive content, including partial nudity and sexual humor.*
> *•Violence, including display of serious injury and events related to violent extremism.*
> *•Inappropriate language, including harassment, profanity and vulgar language.*
> *•Promotion of drugs and regulated substances, including selling, use and abuse of such items.*
> *•Controversial or sensitive subjects and events, including subjects related to war, political conflicts, natural disasters and tragedies, even if graphic imagery is not shown."*
>
> Google support, YouTube, 11th January 2017 [6]

Prior to being publicly available on the platform, videos are processed by an automated moderation system to assess their suitability for monetisation. The platform analyses published videos

[4] Hollister, S. (2012, April 13). YouTube opens Partner program to all: every creator in 20 countries can now monetize video. *The Verge*. https://www.theverge.com/2012/4/13/2945243/youtube-partner-program-monetization

[5] Archive version of the critera page for YouTube partnership was accessed through Internet archives initiative "WayBack Machine": https://web.archive.org/web/20150722172422/https://support.google.com/youtube/answer/82839

[6] Archive version of the guidelines was accessed through Internet archives initiative "WayBack Machine": https://web.archive.org/web/20170111011202/https://support.google.com/youtube/answer/6162278?hl=en

and their related information, such as title, description, and tags [7]. Creators receive a notification if their content is "*flagged as inappropriate for advertising*". Besides algorithmic moderation, the platform also relies on the viewers' and advertisers' feedback to assess videos' suitability for monetisation.

1.2. YouTube's first Adpocalypse

YouTube's Adpocalypse refers to the events in which advertisers pulled advertisements from the platform. These actions were done in reaction to ads being placed next to controversial videos [8]. These withdrawals led YouTube to announce and implement changes in its content moderation policies (Kumar, 2019). In this paper, we examine creators' reactions to YouTube's first moderation policy change that occurred in March 2017. Table 1 describes the timeline of the events.

Table 1: Timeline of events occurred in early 2017 regarding advertisers' massive departures from YouTube.

Date	**Event**
9th February 2017	The Times released an investigation on Google ads placement next to controversial videos on YouTube.
16th March 2017	The Guardian and other brands started to pull off their ads from YouTube.
20th March 2017	YouTube released a statement presenting the actions to reinforce moderation on the platform.
7th April 2017	YouTube announced new criteria for application to the YouTube's Partner Program (criteria of lifetime views).
1st June 2017	YouTube updated its advertiser-friendly guidelines for creators.

In February 2017, the news media "*The Times*" released an investigation on Google's ad placement next to violent, controversial videos [9]. Major brands and advertisers decided to remove their advertisements from YouTube, contesting its lack of control of published content [10]. To retain brands and their revenues, YouTube introduced greater control for advertisers on the placement of their ads across available content [11]. In addition, the platform stated to have reinforced automated

[7] Advertiser-friendly content guidelines – YouTube Help, Google Support: https://support.google.com/youtube/answer/6162278?hl=en

[8] Alexander, J. (2018, May 10). The Yellow $: a comprehensive history of demonetization and YouTube's war with creators. *Polygon*. https://www.polygon.com/2018/5/10/17268102/youtube-demonetization-pewdiepie-logan-paul-casey-neistat-philip-defranco

[9] Mostrous, A. (2017, February 9). Google faces questions over videos on YouTube. *The Times*. https://www.thetimes.com/business-money/technology/article/google-faces-questions-over-videos-on-youtube-3km257v8d

[10] Martinson, J. (2017, March 16). Guardian pulls ads from Google after they were placed next to extremist material. *The Guardian*. https://www.theguardian.com/media/2017/mar/16/guardian-pulls-ads-google-placed-extremist-material

[11] Bardin, A. (2017, March 20). Strengthening YouTube for advertisers and creators. *blog.youtube*. https://blog.youtube/news-and-events/strengthening-youtube-for-advertisers/

(algorithmic) content moderation around videos that violated community guidelines, i.e., to have widened demonetisation criteria. This was followed by a change in eligibility criteria for the partnership program in April 2017 [12]. These revisions were introduced *"to protect brands and creators*" and "*to strengthen advertiser confidence*" [13]. However, these modifications primarily affected creators' monetisation opportunities as the latter are subject to two eligibility criteria: being a member of its Partnership Program (YPP) and following its community guidelines.

The expansion of the demonetisation criteria has some implications for cultural production. This moderation adjustment targeted uploaded content on the platform and incentivised creators to publish advertiser-friendly content. Therefore, by implementing these new guidelines, YouTube sought to influence the production of content and curate the flux of uploaded videos (Grimmelmann, 2015).

Section 2: Prior literature

2.1. Platform governance and content moderation

Platforms may act as regulators and rely on diverse instruments to shape the interactions (Boudreau & Hagiu, 2009). Among these tools, platforms may set rules and practices that elucidate the types of interactions and content allowed on their platforms. It also imposes sanctions for violations of those guidelines. This policing action is referred to as moderation. It is part of a large set of tools coined as platform governance. These normative mechanisms can be implemented after the introduction of new state-level regulation (e.g., NetzDG in Germany, the Digital Service Act in the European Union) or in response to advertisers' pressure to create a brand-safe environment (Griffin, 2023).

Moderation is a complex, multi-layered concept. It may target different aspects of users' participation on the platform, affecting either users' access or uploaded content (Jiang et al., 2023). Platforms may remove, filter and suspend users and content to "*discipline content creation*" (Zeng & Kaye, 2022). Much of the literature has focused its attention on examining the efficiency of moderation measures to reduce the volume of harmful, toxic content on platforms (Aridor et al., 2024). In that scenario, platforms tend to remove rule-breaking content (Ribeiro et al., 2023; Srinivasan et al., 2019) or to deplatform the users violating the guidelines (Jhaver et al., 2021).

[12] Bardin, A. (2017, April 6). Introducing expanded YouTube partner program Safeguards to protect creators. *blog.youtube*. https://blog.youtube/news-and-events/introducing-expanded-youtube-partner/

[13] Schindler, P. (2017, March 21). Expanded safeguards for advertisers. *Blog.google*. https://blog.google/technology/ads/expanded-safeguards-for-advertisers/

Previous research has found that the implementation of these actions leads to a decrease in the volume and the intensity of hateful posts (Andres & Slivko, 2023; Beknazar-Yuzbashev et al., 2024).

Changes in platform governance have some implications for the platform dynamics. The implementation of content moderation policies is motivated by economic incentives such as the maximisation of the platform's profit (Liu et al., 2022; Madio & Quinn, 2024). Moreover, it may affect the sides' levels of participation. For instance, the reduction in the number of available content may lead to a decrease in users' participation and/or in the number of website visits. By examining the market of adult websites, Madio et al. (2025) found evidence that the massive removal of a large portion of videos leads to a long-term decrease in a website's traffic. Moreover, the removal of a user's published comments may lead to a decrease in his activity (Ribeiro et al., 2023). However, this effect is temporary and users gradually returned to their previous level of activity.

2.2. Platform partnership status and users' participation

Platforms can also implement incentives and rewards to shape users' participation and the production of user-generated content. For instance, Burtch et al. (2022) examined the effect of peer rewards on users' activity on Reddit. They found evidence that this incentive increased the volume of uploaded content and users' participation. The introduction of monetary rewards like ad-revenue-sharing programmes (Sun & Zhu, 2013) or tipping (Geng & Chen, 2018) also leads to an increase in creators' content supply and in its quality. These incentives affect creators' extrinsic motivations. Törhönen et al. (2019) provided evidence that non-monetary and financial rewards are complementary. They underlined the influence of the latter on the amount of time invested in the platform by the creators.

Monetary incentives may also be used to moderate users' activity. Platforms can adjust the eligibility criteria for content monetisation or remove them (Zhang et al., 2024). Previous research has studied the effect of these modifications on content supply, particularly in the context of YouTube's Adpocalypse. Abou El-Komboz et al. (2023) examined the causal effect of the change in the platform's partnership programme requirement following the 2017 advertisers' massive departure. Their study describes a decrease in the volume of uploads from creators who lost access to the program. They also found evidence of a decrease in the quality and diversity of uploads. In addition, the removal of monetisation opportunities may encourage creators to publish content on other competing platforms. In their work, Andres et al. (2023) show an increase in activity on Patreon from creators who have accounts on both YouTube and Patreon.

2.3. Demonetisation and creators' participation

In addition to adjusting partnership requirements, the platform can introduce rules that affect users' content monetisation opportunities, such as demonetisation. The logic behind this sanction is as follows: the uploaded rule-breaking content is hosted by the platform, but advertisements are not placed next to it. This process allows advertisers not to be matched with potentially harmful content creators. The absence of a legal definition of brand-safe content allows platforms to self-define this notion, categorising which content is suitable for advertisers (Hill, 2025). Nevertheless, the demonetisation decision is opaque to producers, as little information is provided (Ma & Kou, 2021).

Despite advertising revenue being a source of creators' income (Rieder et al., 2023), demonetisation policies have received little attention from scholars. Dunna et al. (2022) studied the causal effect of this moderation decision on creators' future channel growth, measured by the number of views. They found that the demonetisation status of a video leads to a decrease in its number of views compared to monetised ones. As visibility metrics shape creators' strategies (Duffy et al., 2021), a decrease in the number of views may not motivate them to upload content on the platform.

Following this finding, the introduction of a broader demonetisation policy may influence creators' participation on the platform. As other forms of incentives founded on extrinsic motivations influence content supply, we expect that the number of uploaded videos per non-advertiser-friendly channels will decrease after the implementation of the demonetisation policy in March 2017. This study will test the following hypothesis:

Hypothesis 1: A demonetisation policy tends to decrease the volume of video supply from non-advertiser-friendly content creators.

Content creators' economic models, i.e., the production and monetisation strategies, vary between producers. As the revenue earned from YouTube's ad revenue sharing program depends on the number of views on the published video [14], creators' audience size (number of views and subscribers) has a positive influence on the success of their monetisation strategies (Budzinski & Gaenssle, 2018; Han, 2020). Therefore, the effect of demonetisation may differ per group of producers. This leads to hypothesis two:

[14] Archive version of the "YouTube partner earnings overview" was accessed through Internet archives initiative "WayBack Machine": https://web.archive.org/web/20160505005913/https://support.google.com/youtube/answer/72902

Hypothesis 2: The more subscribers a non-advertiser-friendly content creator has, the more negative is the influence of demonetisation on said creators' participation.

Section 3: The empirical setting

3.1. YouNiverse: YouTube's channels and videos data collection

Collecting data from platforms represents a challenge for researchers. The size of the user base and content collections, and the limited access to platforms' Application Programming Interfaces (APIs) are among the many obstacles to examining activity on social media platforms (Bruns, 2019). To get around these barriers, scholars have made publicly available collections of data retrieved with diverse sampling strategies.

For this research, we use data from the *YouNiverse* collection (Ribeiro & West, 2021). It provides information on more than 136,000 channels and more than 72 million videos published on YouTube between 2005 and 2019. The data were collected in late 2019 from three distinct sources: ChannelCrawler, SocialBlade and YouTube's API. This dataset was selected because it provides data for videos published at the time of the March 2017 moderation change on YouTube and later. Other publicly available collections, e.g., Wu et al.'s dataset [15], gathered information retrieved prior to the moderation update in March 2017.

As noted by Bärtl (2018) in his longitudinal study of the platform, it is impossible to collect and analyse the data of all the content available on YouTube. However, this issue can be overcome by creating a sample that is representative of all the platform. The editors of the *YouNiverse* dataset estimated that it "*covers around 25% of the top 100k most popular YouTube channels and around 35% of the top 10k most popular (both measured by number of subscribers)*" (Hua et al., 2022).3.2. Data

From this database, a sample of channels based on four distinct criteria was created: channels' creation date, channels' activity, channels' category and availability of channels' information in YouNiverse's timeseries dataset [16]. Channels created before the 1st of January 2016 were selected. Channels' activity refers to the presence of videos being uploaded before March 2017. We exclude channels which have not published content prior to YouTube's first moderation update. The channels'

[15] Wu S, Rizoiu M-A, Xie L (2018) Beyond Views: Measuring and Predicting Engagement in Online Videos. Proceedings of the International AAAI Conference on Web and Social Media 12. https://doi.org/10.1609/icwsm.v12i1.15031

[16] This criterion enables grouping selected channels per average number of subscribers and estimating for each group the weekly video supply per channel.

category is defined as the main video category of content [17] uploaded between August 2016 and March 2017. Sixteen categories are referenced in the dataset. As this research focuses on social media native content creators, we exclude channels whose videos are mainly associated with traditional media (films and animation, music, news and politics, and sports). Table 2 presents the channel distribution per main category. Finally, we exclude channels whose weekly volume of content supply may define them as outliers, i.e., channels that posted more than 38 videos per week [18].

From these sampled channels, we retrieved videos uploaded between August 2016 and October 2017 [19]. As the *YouNiverse* dataset provides information at the video level, we constructed a panel at the channel and week level for the observed periods. The final sample is composed of 51,859 channels and 5,596,628 videos.

Table 2: Channels distribution per main category of videos published before the moderation update.

Content Category	**Type of content**	**Content creators count**	**Percentage**
Autos & Vehicles	Content related to auto-mobiles and their technology	2,306	4.4%
Comedy	Comedy content	2,094	4%
Education	Educational content: explanation	3,978	7.7%
Entertainment	Mixture of different genres	10,007	19.3%
Gaming	Content related to video games	10,832	20.9%
How to & Style	Tutorials on various topics	7,498	14.5%
Non-profits & Activism	Content related to politics	664	1.3%
People & Blogs	Personal blogs	8,231	15.9%
Pets & Animals	Videos about / with animals	658	1.3%
Science & Technology	Content related to sciences and technologies: explanation, reviews	2,605	5%
Travel & Events	Content related to travels	1,092	2.1%
Mixed content	Mixture of different categories	1,894	3.7%
Total		51,859	100%

[17] YouTube's API and YouNiverse dataset provide the category for each uploaded video. Budzinski and Gaenssle (2018) defined each video category. For this research, we follow the same definition.

[18] Excluded channels and their uploads represented 0.5% of uploads and 3.1% of sampled channels.

[19] As we examine the causal effect of YouTube's first moderation update, the data collection stops before the events of the second "Adpocalypse". The latter started in November 2017 with a news article from The New-York Times and The Times reporting YouTube's mismanagement around videos for children. This allows for making a distinction in the effect of the two events.

3.3. YouTube's enforcement of content moderation

In this paper, we analyse the causal effect of the enforcement of a "*broader demonetisation*" policy that occurred in late March 2017. Although the new moderation policy was deployed for all YouTube content creators, it targeted specific videos: "*videos that are perceived to be hateful or inflammatory*" and "*content that is harassing or attacking people based on their race, religion, gender or similar categories*" [20]. This description of content is broad and may encapsulate several elements of the advertiser-friendly guidelines.

To estimate the effect on creators' content supply, we categorise channels based on the topic of their published videos before the implementation of the moderation. The categorisation relies on keyword matching. Keywords are used to provide information (metadata) about the content to platforms' algorithmic management systems and users (Firoozeh et al., 2020). As described earlier, the YouTube moderation system analyses this information to classify uploaded videos. Prior to matching, we created a list of keywords related to YouTube's initial non-advertiser-friendly criteria. A list of 96 words and expressions based on this condition was generated (the list of keywords is provided in Appendix A). For each category of criteria, the list follows the informational [21] and univocity properties of keywords.

The keyword matching is done on the title, description and tags for videos published before the March 2017's moderation update (n_{videos} = 2,742,395). The observation period is from August 2016 (calendar week 33) to March 2017 (calendar week 11). For each channel, we calculate the percentage of videos which contain non-advertiser-friendly keywords. This percentage allows the identification of channels that might be affected by the enforcement of the new moderation rule. Creators who never used sensitive keywords in their upload metadata before the new policy implementation are gathered in the advertiser-friendly group (also labelled as the brand-safe one). Others are considered to be affected by the new rule. They will be designated as exposed ones, non-advertiser-friendly ones, or not-suitable-for-advertisers ones. For instance, if a channel has at least one video with a non-advertiser-friendly keyword, it will be labelled as a channel exposed to the new policy.

As a result, 39% of channels in the sample were labelled as non-advertiser-friendly ones ($n_{non\text{-}advertiser\text{-}friendly}$ = 20,323) and 61% as brand-safe ones ($n_{brand\text{-}safe}$ = 31,536). Table 3 presents the

[20] Bardin, A. (2017, March 20). Strengthening YouTube for advertisers and creators. *blog.youtube*. https://blog.youtube/news-and-events/strengthening-youtube-for-advertisers/

[21] Informational properties refer to the principles of exhaustivity, specificity, minimality, impartiality and representativity.

distribution of channels per number of subscribers and per their upload characteristics.Table 3: Channels distribution per audience's size and uploads type

	Advertiser-friendly channels	Non-advertiser-friendly channels	Total
Less than 10,000 subscribers	15,126	8,343	23,469
Between 10,000 and 100,000 subscribers	12,431	8,102	20,533
Between 100,000 and 1,000,000 subscribers	3,499	3,278	6,777
More than 1,000,000 subscribers	480	600	1,080
Total	31,536	20,323	51,859

Section 4: The econometric model

4.1. Model

To estimate the causal effect of the adjustments of the community guidelines, we use a differences-in-difference approach. We compare the volume of content supply of non-advertiser-friendly channels to unaffected channels before and after the implementation of YouTube's broader demonetisation policy. The econometric model is as follows:

$$y_{cw} = \delta(Naf_c * Mod_w) + \mu_c + \rho_w + \varepsilon_{cw}$$

The dependent variable y_{cw} refers to the content supply of channel *c* at week *w*. It is observed for 51,859 creators and for a period of 63 weeks from August 2016 to October 2017. In addition to the outcome variable, this empirical model requires two indicators. Naf_c is a dummy variable indicating whether the channel *c* is labelled as non-advertiser-friendly one. Mod_w is a binary variable indicating the week's position, whether the week *w* is after YouTube's moderation update. The week when YouTube's announced new moderation rules and the ones following the new policy implementation are defined as *after*. Table 4 displays the statistical summary for these indicators.

Under the parallel trend assumption, the coefficient δ measures the causal effect of the moderation policy change on content supply of exposed channels compared to unaffected ones (Callaway, 2022). μ_c and ρ_w respectively represent the channel and time fixed effect *"that account for both channel-specific and time-specific unobserved confounders"* (Imai & Kim, 2021).

Table 4: Statistical summary for variables in the estimated equation.

Variables	Count	Mean	Median	Standard Deviation	Minimum	Maximum
Weekly uploads per channel	3,267,117	1.71	1	3.07	0	38
Non-advertiser-friendly channels	3,267,117	0.39	0	0.49	0	1
Post-moderation change	3,267,117	0.51	1	0.5	0	1

Given that 46% of the channel-week level observations of this study are equal to 0 (Table 5 in Appendix C) [22], the estimation of the average proportional treatment effect should be estimated with caution. The presence of this large proportion of observation can bias the estimated marginal or average marginal effect (Mullahy & Norton, 2024). In order to get reliable estimates of the policy parameters, common practice is to transform the dependent variable in $\ln(y_{cw} + 1)$. This transformation allows to estimate the average proportional demonetization policy effect (average proportional treatment effect $\theta_{ATT\%}$) which is the *"percentage of change in the average outcome for the treated group in the post-treatment period"* (Chen & Roth, 2024). Further explanations are provided in Appendix D. For each analysis, the y_{cw} and $\ln(y_{cw} + 1)$ specifications are estimated.

In order to assess the influence of YouTube's new broader demonetisation rules, we estimate the equation for each group of content creators based on their audience size, i.e. their number of subscribers. Table 5 presents the statistical summary of the outcome variable for each group of content creators.

Table 5: Statistical summary for weekly uploads per channels, week and channels' size.

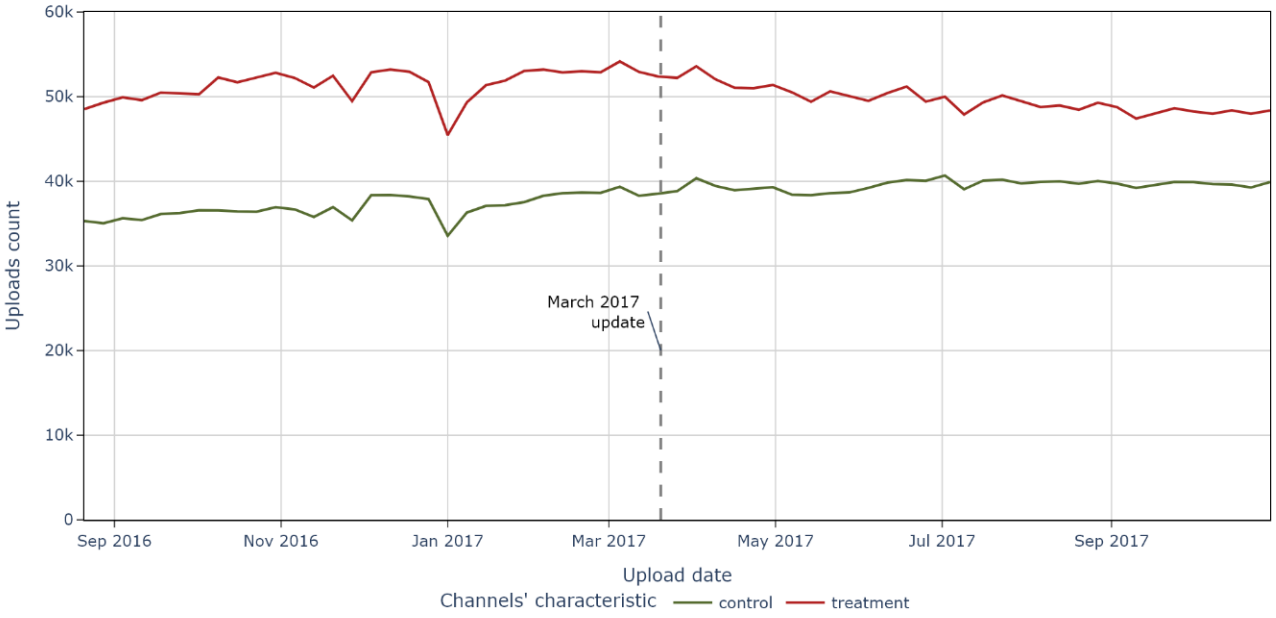
Weekly uploads per channel	Count	Mean	Median	S.D.	Min.	Max.
Less than 10,000 subscribers	1,478,547	1.53	1	2.82	0	38
Between 10,000 and 100,000 subscribers	1,293,579	1.66	1	2.95	0	38
Between 100,000 and 1,000,000 subscribers	426,951	2.25	1	3.43	0	38
More than 1,000,000 subscribers	68,040	3.3	2	4.24	0	38

[22] This can be ascribed by channels' publication frequency: some creators in the sample do not post videos every week.

4.2. Testing for pre-trends

In this section, we check for the presence of potential trends in the weekly video supply before the implementation of a new moderation policy through an event-study approach. Figure 1 illustrates the evolution of weekly video supply of brand-safe creators and non-advertiser-friendly ones between August 2016 and October 2017. Prior to the introduction of the new rules in March 2017, channels in both groups follow the same trend. A similar trend is observed per channel's audience size (Appendix C).

Figure 1: Evolution of the weekly video supply per channels' content type



To confirm this result, we estimate the following model:

$$y_{cw} = \mu_c + \rho_w + \sum_{p=-31;p\neq 0}^{p=31} \beta_w (Naf_c * p) + \varepsilon_{cw}$$

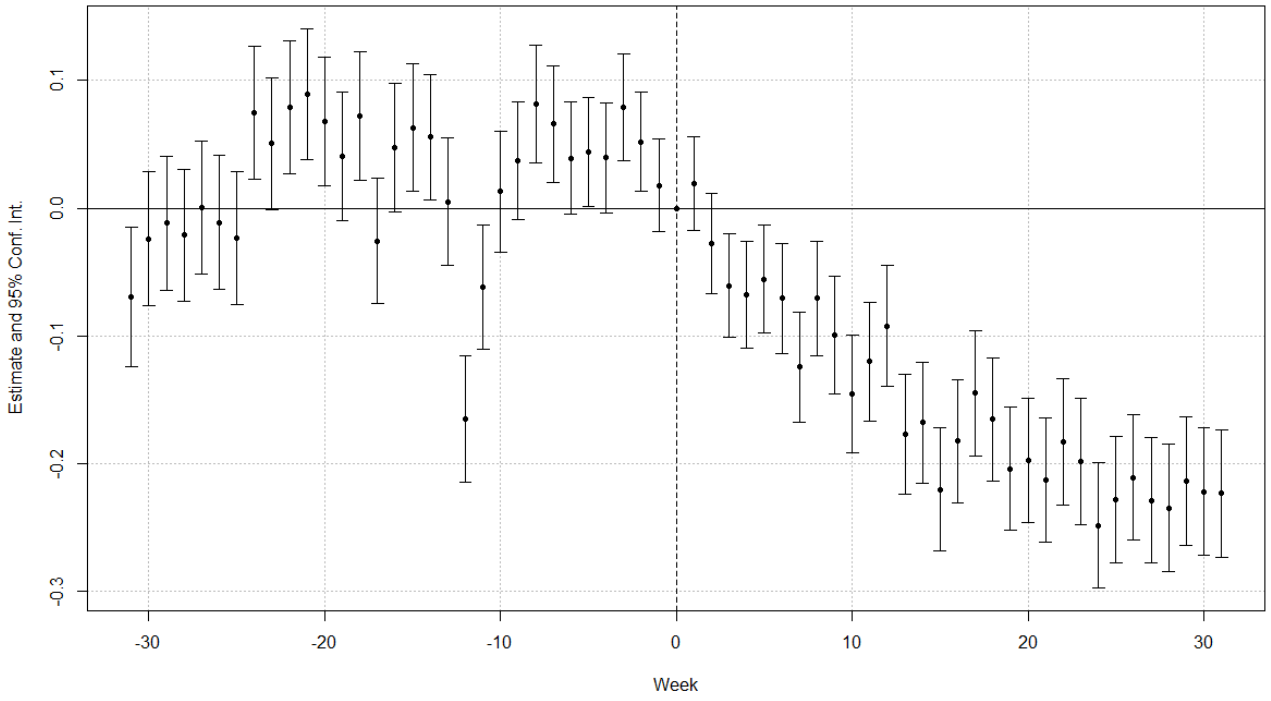
where y_{cw} is the outcome variable. The dummy variable Naf_c is equal to 1 if the channel is considered as non-advertiser-friendly. The variable p refers to the observed week position which is the position according to its distance to the week when YouTube announced its new policy. The period 0 is the week of the announcement of YouTube's moderation policy week (calendar week 12, from 20th to 26th March 20217). The variable p takes values between -31 (first observed week in our sample) and 31 (last observed week). μ_c and ρ_w are the channel and time fixed effects.

Figure 2 provides a visual representation of the result estimation of the difference between non-advertiser-friendly and brand-safe content creators for each week. For the pre-policy change period,

we observe that non-advertiser-friendly creators published more videos per week compared to brand-safe producers. However, this difference is not statistically significant for most of the periods and the estimated coefficients are close to 0 (more details in Appendix C).

Furthermore, for the post-policy change period, the estimated coefficient is statistically significant for each observed week. Figures 1 and 2 show that after the implementation of the demonetisation policy, the weekly volume of videos published by non-advertiser-friendly producers drops compared to brand-safe creators. YouTube's implementation of their demonetisation policy represents a shock in in non-advertiser-friendly channels' content supply. Therefore, we cannot reject the hypothesis that the parallel trend assumption was satisfied in the pre-moderation change period.

Figure 2: Estimated coefficient plot.



Section 5: Results

5.1. The evolution of the weekly content supply

5.1.A. The evolution of weekly uploads per channel

Table 6 displays the results of the model's estimations for all the sampled content creators. It shows a statistically significant reduction in the volume of weekly content supply of non-advertiser-friendly channels compared to brand-safe ones at a 95% confidence level. Similar results are found when we apply a log transformation to the outcome variable $\ln(y_{cw} + 1)$ (Column 2).

In addition, to estimate the marginal effect, we calculate the average proportional treatment effect on exposed channels ($\theta_{ATT\%}$). After the implementation of the new moderation policy, the weekly volume of content supply from exposed creators decreased on average by 10% compared to non-exposed ones. These provide evidence that the enforcement of a demonetisation rule had a deteriorating effect on creators' behaviour (hypothesis H1).

Table 6: Econometric regressions estimations results

	(1) Weekly content supply per channel (y_{cw})	(2) ln ($y_{cw}+1$)
Non-advertiser-friendly channels * post-moderation change	-0.171*** (0.012)	-0.063*** (0.003)
Channel fixed effect	Yes	Yes
Time fixed effect	Yes	Yes
Standard Error clustered by	Channel	Channel
Observations count	3,267,117	3,267,117
R^2	0.615	0.59
Within R^2	0.0005	0.0011
Average proportional demonetization policy effect on non-advertiser-friendly channels ($\theta_{ATT\%}$)	-10%	-10%

** p<0.1; ** p<0.05; *** p<0.001*

Standard error clustered by channel in parenthesis.

5.1.B. The evolution of weekly uploads per channel's main content category

The previous paragraph provides insights into the demonetisation effect on non-advertiser-friendly YouTube channels. However, there is a need to consider whether the weekly content supply volume varies among channels' main content categories. Table 7 displays the estimation results for the three most popular channel categories (gaming, entertainment and people & blogs) compared to the other ones in the study's sample.

We find that there is a statistically significant influence of the demonetisation policy on the content supply of only one group of producers compared to others: those who produce non-advertiser-friendly and gaming-related content. This result may be explained by the use of terms related to war and shooting-related video games. YouTube's demonetisation policy's objective is to reduce the volume of harmful and violent content on the platform. To do so, YouTube relies on automated moderation tools that classify content based on the keywords in videos' metadata. However, these technologies do not take into account the context of the video (Gorwa et al., 2020). Some videos may be labelled as non-advertiser-friendly by mistake. Therefore, the misclassification of a video's monetisation status may discourage creators from being active on the platform.

Table 7: Results of econometric regressions estimations per channel's characteristics.

	(1) Weekly content supply per channel (y_{cw})	(2) ln ($y_{cw}+1$)
Non-advertiser-friendly channels * post-moderation change	-0.132*** (0.0155)	-0.050*** (0.004)
Post-moderation change * Channels with main category gaming	-0.035 (0.024)	-0.007 (0.006)
Post-moderation change * Channels with main category entertainment	-0.016 (0.016)	-0.007* (0.004)
Post-moderation change * Channels with main category People & Blogs	-0.017 (0.017)	0.002 (0.005)
Non-advertiser-friendly channels * post-moderation change * Channels with main category Gaming	-0.128*** (0.038)	-0.031*** (0.009)
Non-advertiser-friendly channels * post-moderation change * Channels with main category Entertainment	-0.005 (0.034)	-0.010 (0.008)
Non-advertiser-friendly channels * post-moderation change * Channels with main category People & Blogs	0.011 (0.035)	-0.008 (0.009)
Channel fixed effect	Yes	Yes
Time fixed effect	Yes	Yes
Standard Error clustered by	Channel	Channel
Observations count	3,267,117	3,267,117
R^2	0.621	0.594
Within R^2	0.0007	0.0012

** p<0.1; ** p<0.05; *** p<0.001*

Standard error clustered by channel in parenthesis.

5.1.C. The evolution of weekly uploads per number of subscribers

In addition to testing for treatment heterogeneity across channel categories, we estimated the causal effect of the demonetisation policy on the content supply per size of creators' audience. To do so, we calculated the average audience size for each channel, i.e., their average subscriber count, between August 2016 and March 2017. Then, we regrouped content creators into four distinct groups.

Table 8 presents the results of the estimation for each group. First, it reveals a statistically significance decrease in the weekly volume of videos for each group of non-advertiser-friendly content creators compared to advertiser-friendly ones. Second, the intensity of the effect varies from one group to another. The treatment effect on exposed channels decreases as a creator's audience size increases. We found that non-advertiser-friendly content creators with less than 10,000 followers uploaded 12% fewer videos per week than their brand-safe colleagues. It is the greatest average treatment effect on exposed channels ($\theta_{ATT\%}$) among the four groups per audience size. This finding

is consistent with the work of Abou El-Komboz et al. (2023). A change in the economic incentives led to a decrease in participation of content creators with a smaller audience on the platform. Moreover, the estimation of the model reveals that brand-unsafe creators with more than a million subscribers published 1% less content per week than brand-safe producers. It is important to notice that the demonetisation effect for this group of content creators is no longer statistically significant when compared to the one for non-advertiser-friendly channels with less than 10,000 subscribers (Appendix E).

This result reveals the following trend regarding the effect of demonetisation policy compared to creators' audience size: the more subscribers a non-advertiser-friendly content creator has, the smaller the treatment effect on weekly video supply will be. These results do not support hypothesis H2. The difference in the average treatment effect could be explained by the diversity in creators' monetisation strategies. Some video producers may not primarily rely on the ad-revenue-sharing programme and have other external sources of income (Glatt, 2022; Hua et al., 2022; Rieder et al., 2023; Alexandre et al., 2024). Furthermore, this reveals the limit of the effectiveness of a uniform demonetisation policy as a moderation tool for platforms.

Table 8: Results of econometric regressions estimations per channel subscribers' count.

	Less than 10,000 subscribers		**Between 10,000 and 100,000 subscribers**		**Between 100,000 and 1,000,000 subscribers**		**More than 1,000,000 subscribers**	
	(1) Weekly content supply per channel (y_{cw})	(2) $\ln(y_{cw}+1)$	(3) Weekly content supply per channel (y_{cw})	(4) $\ln(y_{cw}+1)$	(5) Weekly content supply per channel (y_{cw})	(6) $\ln(y_{cw}+1)$	(7) Weekly content supply per channel (y_{cw})	(8) $\ln(y_{cw}+1)$
Non-advertiser-friendly channels * post-moderation change	-0.098*** (0.020)	-0.050*** (0.004)	-0.171*** (0.019)	-0.058*** (0.004)	-0.24*** (0.022)	-0.067*** (0.005)	-0.17*** (0.033)	-0.037*** (0.006)
Channel fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Standard errors clustered by	Channel	Channel	Channel	Channel	Channel	Channel	Channel	Channel
Observation counts	1,478,547	1,478,547	1,293,579	1,293,579	426 951	426 951	68 040	68 040
R^2	0.543	0.524	0.614	0.589	0.72	0.68	0.8	0.78
Within R^2	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Average proportional demonetisation policy effect on non-advertiser-friendly channels $\theta_{ATT\%}$)	-12%		-8%		-5%		-1%	

** p<0.1; ** p<0.05; *** p<0.001*

Standard error clustered by channel in parenthesis.

5.2 The impact of content moderation on non-advertiser-friendly content

As demonstrated in the previous sub-section, the implementation of a broader demonetisation policy leads to a decrease in non-advertiser-friendly creators' content supply compared to brand-safe ones. In this section, we explore the two mechanisms complementary to the reduction of supply: the content removal and the reduction of non-advertiser-friendly content.

5.2.A. Content removal

Content removal is a common moderation practice on social media platforms. In their work on unavailable videos on YouTube, Kurdi et al. (2021) found that 10.4% of content is deleted in the first week after its upload. Distinct reasons justify the decision to delete content from the platform. Content may be removed by the platform for infringement of the platform's terms of service. Creators may delete videos as they regret uploading them (Wang et al., 2011). The control and protection of users' online public image may also motivate the removal of the content (Acquisti & Gross, 2006).

Unfortunately, social media platforms like YouTube do not provide information about deleted content. Therefore, researchers need to create their own dataset of content and track the evolution of available information. This study examines data available and collected in late 2019. The time distance between the studied period and the data-retrieving date does not allow us to observe video removal from YouTube.

5.2.B. The effect of the policy on creators' toxicity

As mentioned earlier, YouTube's new advertiser-friendly guidelines broaden the demonetised topic list. However, the platform did not disclose the length of the list nor the manner in which content is classified. This opacity led to a new form of precarity in content production (Duffy et al., 2021). Creators tried to adapt to the new guidelines by theorising and applying new production strategies (Ma et al., 2023). In this section, we investigate how exposed content creators react to this new policy; more precisely, we examine the causal effect of the demonetisation policy on the nature of published content.

To do so, we study the evolution of the share of non-advertiser-friendly content in the weekly supply by exposed creators, $(\frac{number\ of\ videos\ that\ are\ non-advertiser-friendly}{number\ of\ published\ videos} * 100)$. Figure 3 displays the evolution of this rate during the analysed period. We can observe a slight increase in the share of not-suitable-for-advertiser content in the weekly supply before the implementation of new moderation rules. After that, it decreased. Table 9 shows a similar trend: the

average share of non-advertiser-friendly content in weekly video supply decreases by 1.4 percentage points between the two periods. The largest drop is observed for channels with an audience size between 100,000 and 1,000,00 followers, while the smallest one is observed for channels with more than a million subscribers. This would suggest that there is also a variation in the causal effect between content creators.

Figure 3: Evolution of the share of non-advertiser-friendly content in weekly uploads of video for channels impacted by the demonetisation policy.

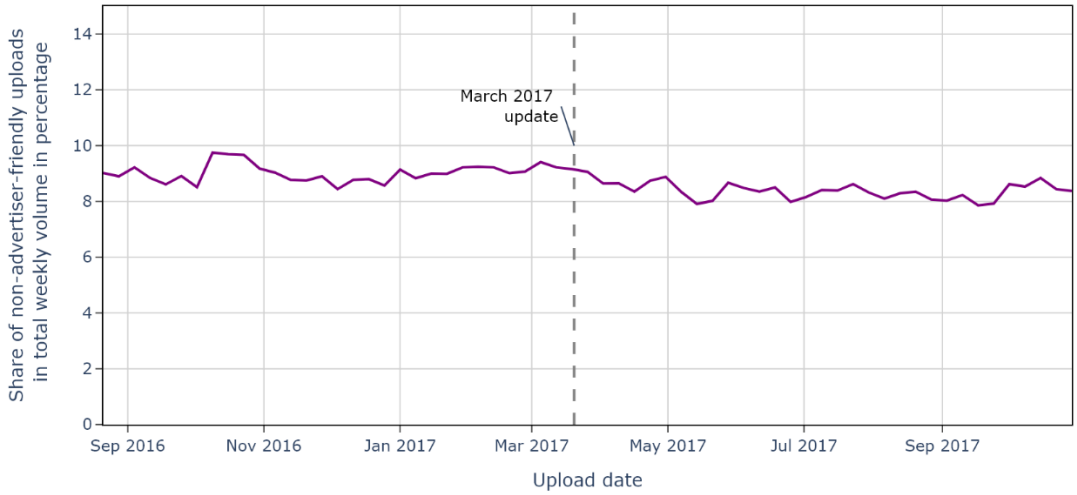


Table 9: Average share of non-advertiser-friendly content in weekly video supply of exposed channels per period.

Share of non-advertiser-friendly content in weekly video supply	Whole sample	Less than 10,000 subscribers	Between 10,000 and 100,000 subscribers	Between 100,000 and 1,000,000 subscribers	More than 1,000,000 subscribers
Before the implementation of new moderation rules	9.97	9.69	10.4	9.86	8.69
After the implementation of new moderation rules	8.57	8.26	9.12	8.21	7.79

Complementary to the statistical description, we use an ordinary least squares (OLS) approach to examine the influence of the new platform policy on the share of non-advertiser-friendly content in weekly video supply. The model is as follows:

$$y_{cw} = Mod_w + \varepsilon_{cw}$$

with y_{cw} referring to the proportion of harmful content in creators' supply and $Mod_w$to the week's position. The latter variable is coded 1 if the observation is after YouTube's enforcement of the demonetisation rule. A statistical summary of the dependent variable is provided in Table 10.

Table 10: Statistical summary of the outcome variable per group of content creators

Non-advertiser-friendly content share	**Count**	**Mean**	**Median**	**S.D.**	**Min.**	**Max.**
Whole sample (exposed channels)	878,806	9.27	0	24.5	0	100
Less than 10,000 subscribers	340,598	8.96	0	24.28	0	100
Between 10,000 and 100,000 subscribers	347,738	9.77	0	25.36	0	100
Between 100,000 and 1,000,000 subscribers	158,505	9.05	0	23.68	0	100
More than 1,000,000 subscribers	31,965	8.25	0	20.86	0	100

Table 11 presents the results of this estimation for the sample of exposed creators and for each group per audience size. The introduction of the demoderation policy led to a statistically significant decrease of the share of non-advertiser-friendly content in the weekly content supply. The share of not-suitable-for advertisers in the volume of content supply per week decreased on average by 1.4 percentage points after the enforcement of the new policy. This result is consistent with previous results found in the literature.

Table 11: Estimations results of the evolution of the share of non-advertiser-friendly content in weekly uploads volume.

	(1) Whole sample	(2) Less than 10,000 subscribers	(3) Between 10,000 and 100,000 subscribers	(4) Between 100,000 and 1,000,000 subscribers	(5) More than 1,000,000 subscribers
Post-moderation change	-1.396*** (0.052)	-1.425*** (0.083)	-1.273*** (0.086)	-1.655*** (0.119)	-0.908*** (0.233)
Constant	9.970*** (0.037)	9.690*** (0.060)	10.398*** (0.060)	9.865*** (0.083)	8.695*** (0.164)
Observations	878,806	340,598	347,738	158,505	31,965
R^2	0.001	0.001	0.001	0.001	0.001

Similar to the volume of weekly uploads (1), the implementation of the demonetisation policy led to a statistically significant decrease of the share of non-advertiser-friendly videos for each group per audience size. However, the intensity of the variation varies between groups. The greatest variation is observed for exposed creators with between 100,000 and 1,000,000 subscribers: the ratio decreases by 1.65 percentage points after the implementation of the new rule. The lowest variation is seen for the ones with more than one million subscribers.

These results provide additional insight into the influence of moderation policy on creators' participation on the platform. The introduction of uniform rules has a heterogeneous influence on producers' participation and supplied content. Several reasons may explain this heterogeneity. First, platforms' governance mechanisms may differ from one creator to another (Feher, 2023). Platforms may not apply strict sanctions for superstar content creators as they gather a large

audience[23]. Second, producers may not rely only on ad-revenue sharing programmes and diversify their sources of income (Glatt, 2022; Rieder et al., 2023). Both mechanisms may ease the threat of an economic sanction.

Conclusions

This paper aims to assess the causal effect of the implementation of YouTube's broader demonetisation policy on content supply. This change occurred in late March 2017 in reaction to advertisers' boycott of the platform. In order to retain advertisers, YouTube announced a series of changes, including a reinforcement of automated content moderation and an expansion of its non-advertiser-friendly criteria list. This research examines the evolution of the video provision strategies of 51,583 channels between August 2016 and October 2017. Two aspects of content creators' production were analysed: the number of posted videos per week and the proportion of videos with not-suitable-for-advertisers' keywords in the weekly content supply.

The adjustment of the demonetisation rule led to a statistically significant drop in the number of videos published by non-advertiser-friendly channels compared to brand-safe ones. On close inspection, the decrease is heterogeneous across channels' main video category and their audience size. In addition, the enforcement of this new moderation policy led to a decrease in the share of non-advertiser-friendly videos in the weekly content supply of the exposed channels. This decrease is also heterogeneous across different groups of producers.

This study provides insights into the consequences of the platform governance mechanisms on creators' production. These findings illustrate the efficiency of this sanction as a way to incentivise the production towards brand-safe videos. As YouTube does not own nor produce the content, the platform may indirectly control the flow of uploads by changing the community guidelines. This mechanism provides evidence of the platform's ability to curate the flow of content and to discipline content creation to create a brand-safe environment.

However, demonetisation is an economic sanction leading potentially to a loss of revenue for content producers. As demonstrated by Dunna et al. (2023), the video monetisation status affects its visibility and the channel's future growth. Since the demonetisation decision is vague, it reinforces the creators' perceived precarity of their activity and their incomes. In this scenario, creators face a trade-off between self-expression and the economic viability of their activity.

[23] Grayson, N. (2021, October 15). The truth behind Twitch's leaked 'do not ban' list. *The Washington Post*. https://www.washingtonpost.com/video-games/2021/10/15/twitch-leak-do-not-ban-streamers-tyler1-ricegum/

Moreover, this research sheds new light on the heterogeneous impact of uniform moderation measure on users' participation. The results of content demonetisation vary among channels' audiences (their subscriber count). This reveals differences in creators' production and monetisation strategies. For example, some creators may rely on advertising revenue to build and maintain their activity. This economic dependence may increase producers' compliance with new rules. This can explain the high variation in the volume of videos published by exposed content creators. However, some creators may no longer rely on the platform's advertising revenue. As they have a bigger audience, they have other ways to monetise their activity, sometimes outside the platform. In their investigation of video producers' monetisation strategies, Rieder et al. (2023) have grouped together five major sources of income, four of which are alternative mechanisms to YouTube's own monetisation system. Similarly, Hua et al. (2022) found that 60.6% of channels present in the YouNiverse dataset have alternative monetisation strategies to advertising revenues. By having alternative and external income sources, creators with large audiences may be less incentivised to change their production strategies. Further research should examine the efficiency of sanctions tailored for each creator's audience size.

Finally, these findings provide some preliminary elements characterising the implications of ad-funded business model on cultural production, more precisely on platformised creative work. The massive departure of brands from YouTube put economic pressure on the platform to change its moderation policy, therefore affecting its design and rules. We provide evidence that advertiser-driven policy may harm the participation of content creators. As online content creation is dependent on platforms' features (Duffy et al., 2021), changes in the latter affect creators' activity and revenues. Its impact is stronger for creators who are dependent on advertising revenues to maintain their activity. This paper highlights the economic interplay between advertisers and content creators, and the trade-offs that platform owners face: attracting advertisers or attracting and retaining content creators (Bhargava, 2021; Jain & Qian, 2021).

This paper is limited to the technical aspects that open up perspectives for further research. A note of caution is due regarding the data. The analysis is conducted on data collected as of late 2019. Therefore, the data is a snapshot of YouTube's media landscape as of late 2019 and it does not take into account the different modifications made by creators or platforms (changes in metadata, content deletion). In addition, the data collection does not provide information on videos' monetisation status. This absence could impact channels' categorisation and the evaluation of the causal effect on content production. Further research may examine the effect of the monetisation status of the video on the creators' level of production.

In addition, this research only investigates the causal effect of demonetisation on the quantity of the supply (weekly published video count). Other characteristics in the platform dynamic may also be affected by this type of moderation, such as the number of views (demand side) or the diversity of published content (supply diversity). The former requires additional data on sampled videos. The latter would need to look closely at the video's details and their context. Moreover, it is important not to put aside the impact on the demand side, as both dimensions would provide a better understanding of the governance mechanisms and a better assessment of the impact at the platform level.

References

Abou El-Komboz, L., Kerkhof, A., & Loh, J. (2023). Platform Partnership Programs and Content Supply: Evidence from the Youtube "Adpocalypse". *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4416792

Acquisti, A., & Gross, R. (2006). Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In G. Danezis & P. Golle (Eds), *Privacy Enhancing Technologies* (pp. 36–58). Springer. https://doi.org/10.1007/11957454_3

Alexandre, O., Benbouzid, B., Lelievre, A., & Roudier, B. (2024). Au marché de Youtube: Organisation, revenus et topologie. *Réseaux*, *N° 246*(4), 43–88. https://doi.org/10.3917/res.246.0043

Andres, R., Rossi, M., & Tremblay, M. (2023). YouTube "Adpocalypse": The YouTubers' Journey from Ad-Based to Patron-Based Revenues. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4672028

Andres, R., & Slivko, O. (2023). Combating Online Hate Speech: The Impact of Legislation on Twitter. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4013662

Aridor, G., Jiménez-Durán, R., Levy, R., & Song, L. (2024). The Economics of Social Media. *Journal of Economic Literature*, *62*(4), 1422–1474. https://doi.org/10.1257/jel.20241743

Beknazar-Yuzbashev, G., Jiménez-Durán, R., & Stalinski, M. (2024). A Model of Harmful Yet Engaging Content on Social Media. *AEA Papers and Proceedings*, *114*, 678–683. https://doi.org/10.1257/pandp.20241004

Bhargava, H. K. (2021). The Creator Economy: Managing Ecosystem Supply, Revenue Sharing, and Platform Design. *Management Science*, 1–19. https://doi.org/10.1287/mnsc.2021.4126

Boudreau, K. J., & Hagiu, A. (2009). Platform Rules: Multi-Sided Platforms as Regulators: Platforms, Markets and Innovation. In *Platforms, Markets and Innovation* (p. 29). Edward Elgar Publishing. https://www.elgaronline.com/view/9781848440708.00014.xml

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Budzinski, O., & Gaenssle, S. (2018). The economics of social media (super-)stars: An empirical investigation of stardom and success on YouTube. *Journal of Media Economics*, *31*(3–4), Article 3–4. https://doi.org/10.1080/08997764.2020.1849228

Callaway, B. (2022). Difference-in-Differences for Policy Evaluation. In K. F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics* (pp. 1–61). Springer International Publishing. https://doi.org/10.1007/978-3-319-57365-6_352-1

Chen, J., & Roth, J. (2024). Logs with Zeros? Some Problems and Solutions. *The Quarterly Journal of Economics*, *139*(2), 891–936. https://doi.org/10.1093/qje/qjad054

Choi, J. P., & Jeon, D.-S. (2023). Platform design biases in ad-funded two-sided markets. *The RAND Journal of Economics*, *54*(2), 240–267. https://doi.org/10.1111/1756-2171.12436

Duffy, B. E., Pinch, A., Sannon, S., & Sawey, M. (2021). The Nested Precarities of Creative Labor on Social Media. *Social Media + Society*, *7*(2), 205630512110213. https://doi.org/10.1177/20563051211021368

Dunna, A., Keith, K. A., Zuckerman, E., Vallina-Rodriguez, N., O'Connor, B., & Nithyanand, R. (2022). Paying Attention to the Algorithm Behind the Curtain: Bringing Transparency to YouTube's Demonetization Algorithms. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 318:1-318:31. https://doi.org/10.1145/3555209

Feher, A. (2023). *How to enforce platforms' liability?*

Firoozeh, N., Nazarenko, A., Alizon, F., & Daille, B. (2020). Keyword extraction: Issues and methods. *Natural Language Engineering*, *26*(3), 259–291. https://doi.org/10.1017/S1351324919000457

Geng, R., & Chen, X. (2018). Economics of 'Tipping' Button in Social Media: An Empirical Analysis of Content Monetization. *PACIS 2018*, 15.

Glatt, Z. (2022). *"We're all told not to put our eggs in one basket": Uncertainty, precarity and cross-platform labor in the online video influencer industry*. International Journal of Communication, Special Issue on Media and Uncertainty. https://ijoc.org/index.php/ijoc/article/view/15761

Goanta, C., Yohanis, A., Jaiman, V., & Urovi, V. (2022). Web monetisation. *Internet Policy Review*, *11*, Article 1. https://policyreview.info/glossary/web-monetisation

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Griffin, R. (2023). From brand safety to suitability: Advertisers in platform governance. *Internet Policy Review*, *12*(3). https://policyreview.info/articles/analysis/safety-to-suitability-advertisers-in-platform-governance

Grimmelmann, J. (2015). The Virtues of Moderation. *Yale Journal of Law and Technology*. https://openyls.law.yale.edu/handle/20.500.13051/7798

Han, B. (2020). How do YouTubers make money? A lesson learned from the most subscribed YouTuber channels. *International Journal of Business Information Systems*, *33*(1), Article 1. https://ideas.repec.org/a/ids/ijbisy/v33y2020i1p132-143.html

Hill, S. (2025). Stop hate for profit: Evaluating the mobilisation of advertisers and the advertising industry to regulate content moderation on digital platforms. *Internet Policy Review*, *14*(1). https://doi.org/10.14763/2025.1.1825

Hua, Y., Ribeiro, M. H., West, R., Ristenpart, T., & Naaman, M. (2022). Characterizing Alternative Monetization Strategies on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, *6*, 1–30. https://doi.org/10.1145/3555174

Imai, K., & Kim, I. S. (2021). On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. *Political Analysis*, *29*(3), 405–415. https://doi.org/10.1017/pan.2020.33

Jain, S., & Qian, K. (2021). Compensating Online Content Producers: A Theoretical Analysis. *Management Science*, *67*(11), Article 11. https://doi.org/10.1287/mnsc.2020.3862

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 381:1-381:30. https://doi.org/10.1145/3479525

Jiang, J. A., Nie, P., Brubaker, J. R., & Fiesler, C. (2023). A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction*, *30*(1), 3:1-3:34. https://doi.org/10.1145/3534929

Kerkhof, A. (2024). Advertising and Content Differentiation: Evidence from YouTube. *The Economic Journal*, *134*(663), 2912–2950. https://doi.org/10.1093/ej/ueae043

Kopf, S. (2020). "Rewarding Good Creators": Corporate Social Media Discourse on Monetization Schemes for Content Creators. *Social Media + Society*, *6*(4), Article 4. https://doi.org/10.1177/2056305120969877

Kumar, S. (2019). The algorithmic dance: YouTube's Adpocalypse and the gatekeeping of cultural content on digital platforms. *Internet Policy Review*, *8*(2), Article 2. https://doi.org/10.14763/2019.2.1417

Liu, Y., Yildirim, T. P., & Zhang, Z. J. (2022). Implications of Revenue Models and Technology for Content Moderation Strategies. *Marketing Science*, *41*(4), 663–869. https://doi.org/10.1287/mksc.2022.1361

Ma, R., & Kou, Y. (2021). 'How advertiser-friendly is my video?': YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 429:1-429:25. https://doi.org/10.1145/3479573

Ma, R., You, Y., Gui, X., & Kou, Y. (2023). How Do Users Experience Moderation?: A Systematic Literature Review. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 278:1-278:30. https://doi.org/10.1145/3610069

Madio, L., & Quinn, M. (2024). Content moderation and advertising in social media platforms. *Journal of Economics & Management Strategy*, *n/a*. https://doi.org/10.1111/jems.12602

Mullahy, J., & Norton, E. C. (2024). Why Transform Y? The Pitfalls of Transformed Regressions with a Mass at Zero. *Oxford Bulletin of Economics and Statistics*, *86*(2), 417–447. https://doi.org/10.1111/obes.12583

Ribeiro, M. H., Cheng, J., & West, R. (2023). Automated Content Moderation Increases Adherence to Community Guidelines. *WWW '23: Proceedings of the ACM Web Conference 2023*, 2666–2676. https://doi.org/10.1145/3543507.3583275

Ribeiro, M. H., & West, R. (2021). YouNiverse: Large-Scale Channel and Video Metadata from English-Speaking YouTube. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 1016–1024. https://doi.org/10.1609/icwsm.v15i1.18125

Rieder, B., Borra, E., Coromina, Ò., & Matamoros-Fernández, A. (2023). Making a Living in the Creator Economy: A Large-Scale Study of Linking on YouTube. *Social Media + Society*, *9*(2), 20563051231180628. https://doi.org/10.1177/20563051231180628

Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L., & Tan, C. (2019). Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 163:1-163:21. https://doi.org/10.1145/3359265

Sun, M., & Zhu, F. (2013). Ad Revenue and Content Commercialization: Evidence from Blogs. *Management Science*, *59*(10). https://doi.org/10.1287/mnsc.1120.1704

Tang, Q., Gu, B., & Whinston, A. B. (2012). Content Contribution for Revenue Sharing and Reputation in Social Media: A Dynamic Structural Model. *Journal of Management Information Systems*, *29*(2), Article 2. https://doi.org/10.2753/MIS0742-1222290203

Törhönen, M., Sjöblom, M., Hassan, L., & Hamari, J. (2019). Fame and fortune, or just fun? A study on why people create content on video platforms. *Internet Research*, *30*(1), 165–190. https://doi.org/10.1108/INTR-06-2018-0270

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. (2011). 'I regretted the minute I pressed share': A qualitative study of regrets on Facebook. *Proceedings of the Seventh Symposium on Usable Privacy and Security*, 1–16. https://doi.org/10.1145/2078827.2078841

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287

Zhang, D., Jiang, H., Qiang, M., Zhang, K., & Qiu, L. (2024). Time to Stop? An Empirical Investigation on the Consequences of Canceling Monetary Incentives on a Digital Platform. *Information Systems Research*, isre.2022.0017. https://doi.org/10.1287/isre.2022.0017

Zimmer, F., & Scheibe, K. (2019, January). What Drives Streamers? Users' Characteristics and Motivations on Social Live Streaming Services. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences 2019 (HICSS-52), Grand Wailea, Hawaii. https://doi.org/10.24251/HICSS.2019.306

Appendix

Appendix A: Channel categorisation with content classification

Appendix Table 1: Presentation of keywords per non-advertiser-friendly criteria.

Type of none advertiser friendly topics	Keyword count	Keyword list
Sexually suggestive content, including partial nudity and sexual humor	16	ass, sex, dick, vagina, penis, anal, anus, erotic, boobs, butt, cum, nudity, sexual, porn, xxx, fuck, condom.
Violence, including display of serious injury and events related to violent extremism	21	blood, torture, murder, abuse, rape, kill, dead, mass shooting, execution, kidnapping, slaughter, suicide, victim, violence, violent, weapon, warefare, gun, bomb, terrorism, extremism
Inappropriate language, including harassment, profanity and vulgar language	9	bastard, pussy, dumbass, goddamn, bitch, n-word, shit, idiot, dammit, cunt
Promotion of drugs and regulated substances, including selling, use and abuse of such items	17	drug, acid, weed, boong, cannabis, cbd, cocain, crack, dealer, joint, junky, lsd, marijuana, rehab, stoned, thc, heroin
Controversial or sensitive subjects and events, including subjects related to war, political conflicts, natural disasters and tragedies, even if graphic imagery is not shown	33	war, abortion, accident, Hitler, fascism, AIDS, Al Qaeda, alt right, genocide, assassination, attack, concentration camp, incel, holocaust, homophobia, illegal, incest, ISIS, Israel, Palestine, jewish, Ku Klux Klan, LGBT, nazi, racism, slavery, supremacist, supremacy, transphobia, nuclear weapon, climate change, 9/11, Twin Towers
Total	96	

Appendix Table 2: Presentation of dataset of videos upload before the moderation update.

	Video count	**Percentage**
Videos with missing description	78,607	2.87%
Videos with missing tags	235,058	8.57%
Videos with missing both description and tags	45,071	1.64%
Videos uploaded between week 33 2016 and week 11 2017	2,742,395	100%

Appendix Table 3: Video distribution per presence of non-advertiser-friendly keywords in title, description and tags.

Type of non-advertiser-friendly topics	Count of videos with non-advertiser-friendly keywords	Percentage of video uploaded
Sexually suggestive content, including partial nudity and sexual humor	15,262	0.56%
Violence, including display of serious injury and events related to violent extremism	87,064	3.17%
Inappropriate language, including harassment, profanity and vulgar language	6,611	0.24%
Promotion of drugs and regulated substances, including selling, use and abuse of such items	11,237	0.41%
Controversial or sensitive subjects and events, including subjects related to war, political conflicts, natural disasters and tragedies, even if graphic imagery is not shown	42,056	1.53%
Total uploaded videos before first moderation update	2,742,395	100%

The keyword matching was processed for each category of non-advertiser-friendly content. Some videos can be counted in several categories.

Appendix B: Average proportional treatment effect on exposed channels

Appendix Table 4: Distributions of observations per number of published videos per week

Weekly content supply volume	Observation counts	Percentage of observations
0 video	1,499,813	45.90%
1 video	725,403	22.20%
2 videos	364,164	11.10%
3 videos	199,590	6.10%
4 videos	118,337	3.60%
Between 5 and 10 videos	282,211	8.60%
More than 10 videos	77,599	2.40%

In order to assess the marginal effect of the introduction of the demonetisation rule on the weekly video supply (y_{cw}), we could calculate the average treatment effect (ATE) through a log-like transformation such as ln (y_{cw}). However, our sample is composed by 46% of observations equal to 0 (Table 2.4). This could affect the definition and the interpretation of the average treatment effect (ATE) (Mullahy & Norton, 2024).

To bypass this issue, we compute the average proportional treatment effect on the treated ($\theta_{ATT\%}$). As mentioned in Section 4, this indicator is the percentage change in the average outcome for non-advertiser-friendly channels after the monetisation rule change in March 2017 (Chen and Roth 2024). The calculation is as follows:

$$\theta_{ATT\%} = \frac{E[Y_{cw}(1) \mid Treat_c = 1, Mod_w = 1] - E[Y_{cw}(0) \mid Treat_c = 1, Mod_w = 1]}{E[Y_{cw}(0) \mid Treat_c = 1, Mod_w = 1]}$$

The latter requires to know the counter factual post-treatment average weekly content supply volume for exposed creators $E[Y_{cw}(0) \vee Treat_c = 1, Mod_w = 1]$. It can be calculated through the application of the parallel trend assumption[24] in its ratio version as described below:

$$\frac{E[Y_{cw}(0) \mid Treat_c = 0, Mod_w = 1]}{E[Y_{cw}(0) \mid Treat_c = 0, Mod_w = 0]} = \frac{E[Y_{cw}(0) \mid Treat_c = 1, Mod_w = 1]}{E[Y_{cw}(0) \mid Treat_c = 1, Mod_w = 0]}$$

If YouTube did not implement a new moderation policy, therefore, the average percentage change in the mean outcome for non-advertiser-friendly creators would have been the same as for brand-safe channels. By calculating the average percentage change in the weekly content supply, I

[24] This assumption was tested in Section 4.2.

can deduce the counterfactual post-treatment average outcome. Table 5 and 6 presents the results of this procedure.

Appendix Table 5: Average weekly volume of content supply of brand-safe creators

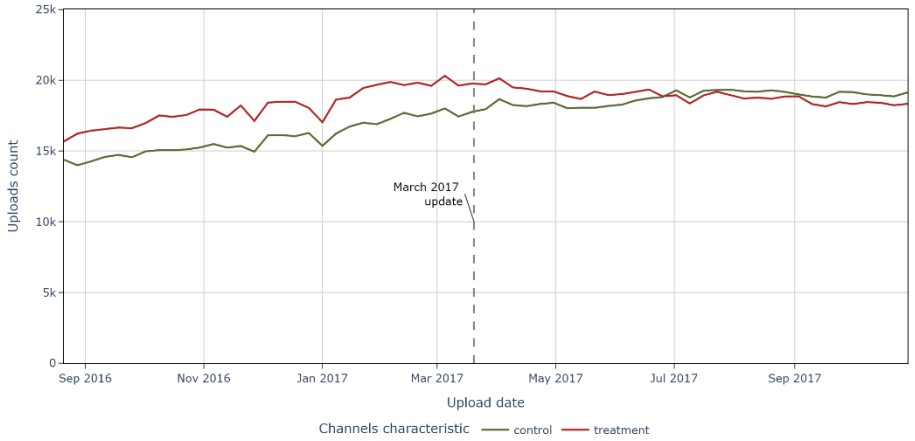
	Whole sample	Less than 10,000 subscribers	Between 10,000 and 100,000 subscribers	Between 100,000 and 1,000,000 subscribers	More than 1,000,000 subscribers
Average weekly supply of unexposed channels before the moderation change	1.17	1.05	1.17	1.55	2.27
Average weekly supply of unexposed channels after the moderation change	1.25	1.24	1.18	1.47	2.11
Variation in percentage	107%	118%	101%	95%	93%

Appendix Table 6: Average weekly volume of content supply of non-advertiser-friendly creators

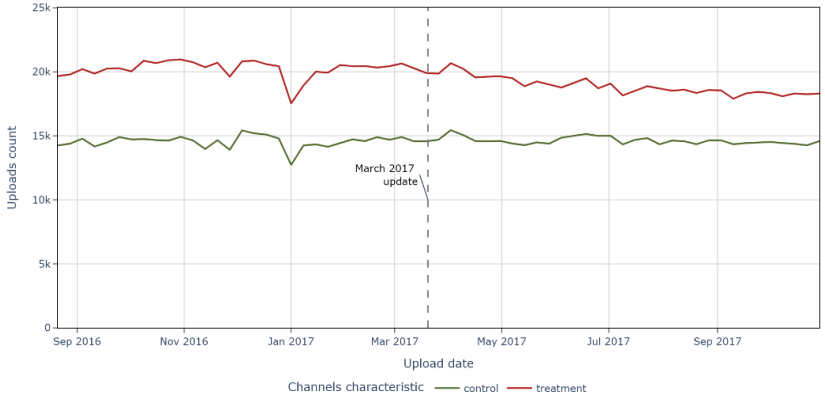
	Whole sample	Less than 10,000 subscribers	Between 10,000 and 100,000 subscribers	Between 100,000 and 1,000,000 subscribers	More than 1,000,000 subscribers
Average weekly supply of exposed channels before the moderation change	2.53	2.17	2.5	3.2	4.36
Counterfactual post-moderation means weekly supply of exposed channels	2.70	2.56	2.52	3.03	4.05
Actual average weekly supply after moderation update	2.44	2.26	2.33	2.89	4.02
Proportional treatment effect	-0.10	-0.12	-0.08	-0.05	-0.01

Appendix C: Weekly supply evolution per channels' size and characteristic

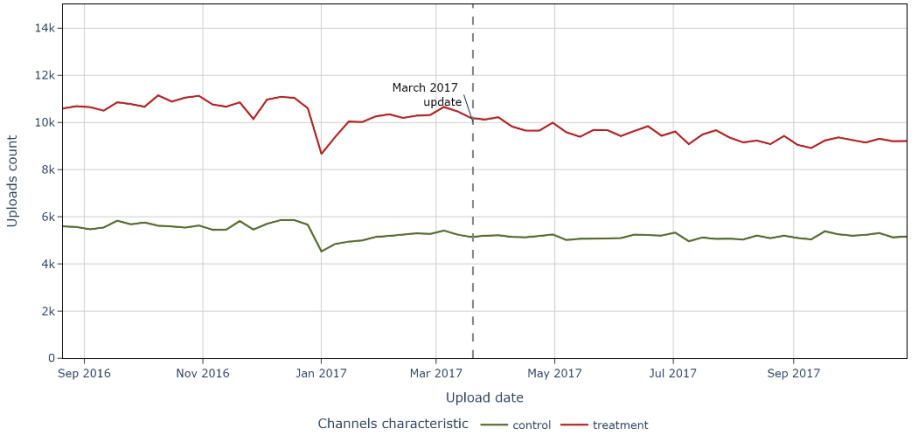
Appendix Figure 1: Evolution of the volume of uploads per channel and per week for creators with on average less than 10,000 subscribers before the moderation update.



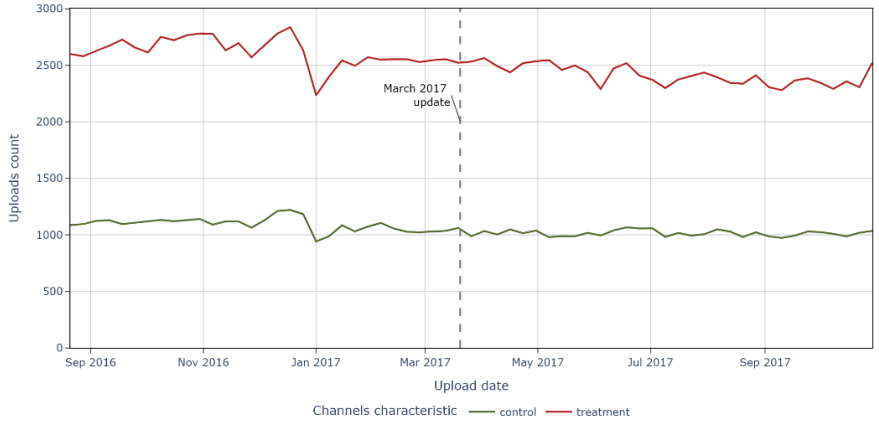
Appendix Figure 2: Evolution of the volume of uploads per channel and per week for creators with on average between 10,000 and 100,000 subscribers before the moderation update.



Appendix Figure 3: Evolution of the volume of uploads per channel and per week for creators with on average between 100,000 and 1,000,000 subscribers before the moderation update.



Appendix Figure 4: Evolution of the volume of uploads per channel and per week for creators with on average more than 1,000,000 subscribers before the moderation update.



Appendix D: Event study regression

Appendix Table 7: Event-study estimations results

	Two-way fixed effect regression
Non-advertiser-friendly channels x Period [-31; -26]	-0.023 (0.023)
Non-advertiser-friendly channels x Period [-25; -21]	0.054** (0.023)
Non-advertiser-friendly channels x Period [-20; -16]	0.041* (0.022)
Non-advertiser-friendly channels x Period [-15; -11]	-0.021 (0.021)
Non-advertiser-friendly channels x Period [-10; -6]	0.047** (0.019)
Non-advertiser-friendly channels x Period [-5; -1]	0.046*** (0.016)
Non-advertiser-friendly channels x Period 0	*Reference*
Non-advertiser-friendly channels x Period [1;5]	-0.039** (0.016)
Non-advertiser-friendly channels x Period [6;10]	-0.102*** (0.019)
Non-advertiser-friendly channels x Period [11;15]	-0.156*** (0.020)
Non-advertiser-friendly channels x Period [16;20]	-0.179*** (0.020)
Non-advertiser-friendly channels x Period [21;25]	-0.214*** (0.022)
Non-advertiser-friendly channels x Period [26;31]	-0.222*** (0.022)
Unit (channel) fixed-effects	Yes
Time (week) fixed-effects	Yes
S.E.: Clustered by	Channel
Observations	3,267,117
R^2	0.615
Within R^2	0.0007

** p<0.1; ** p<0.05; *** p<0.001*

Standard error clustered by channel in parenthesis

Appendix E: Difference-in-differences model with channels' audience size

Appendix Table 8: Results of econometric regression

	Weekly content supply
Non-advertiser-friendly channels * Moderation policy change	-0.098*** (0.020)
Moderation policy * Channels with less than 10k subscribers	
Moderation policy * Channels with between 10k and 100k subscribers	-0.183*** (0.013)
Moderation policy * Channels with between 100k and 1M subscribers	-0.267*** (0.021)
Moderation policy * Channels with more than 1M subscribers	-0.352*** (0.056)
Non-advertiser-friendly channels * Moderation policy change * Channels with less than 10k subscribers	
Non-advertiser-friendly channels * Moderation policy change * Channels with between 10k and 100k subscribers	-0.073*** (0.028)
Non-advertiser-friendly channels * Moderation policy change * Channels with between 100k and 1M subscribers	-0.142*** (0.039)
Non-advertiser-friendly channels * Moderation policy change * Channels with more than 1M subscribers	-0.078 (0.085)
Channel fixed effect	Yes
Time fixed effect	Yes
Standard Error clustered by	Channel
Observations count	3,0267,117
R^2	0.622
Within R^2	0.0018

** p<0.1; ** p<0.05; *** p<0.001*

Standard error clustered by channel in parenthesis