

Enforcers of Truth: Social Media Platforms and Misinformation

Bridget Barrett, Daniel Kreiss, & Madhavi Reddi

Digital Politics Project
UNC Center for Information, Technology
& Public Life

Introduction

Social media companies frequently resist being arbiters of truth. Yet, as their services have been systematically exploited in harmful and manipulative ways, they increasingly carved out categories of content for fact-checking. This report documents how Facebook and Instagram, Reddit, Snapchat, Twitter, and YouTube use both internal teams and third-party companies to fact check content containing manipulated media as well as content related to democratic processes, major tragic events, and health.

To varying degrees, all these social media companies have enacted policies prohibiting false information in at least two of these categories. In this report, we show that platforms largely resist playing the contentious role of arbiters of truth by focusing on problematic content which can be checked against widely accepted and documented truths based in institutional and scientific-consensus. These companies only remove the most uncontroversial and undisputed falsehoods that are most likely to cause harm, rather than taking responsibility for parsing truth from fiction on a broader level. In other words, they attempt to enforce the truth while not being responsible for its arbitration.

The COVID-19 pandemic and respective infodemic have clarified this strategy. In response, researchers, journalists, and public health officials have pointed out the many difficulties in relying on institutionally authoritative sources in a fast-changing health crisis on algorithmically-driven platforms.

The purpose of this report is to provide a side-by-side comparison of social media platforms' differential attempts to enforce truth and some of the considerable challenges they face in doing so. Heading into the 2020 election, it is necessary to document what types of dis- and mis-information related to politics are differentially prohibited on platforms.¹ As researchers investigate what types of misinformation are most common online, who produces it, why people share this content, and the effects this content potentially has, this report documents where major social media platforms are currently willing to step in (and where they are not). We do not address the challenges of enforcement.

We hope this helps to inform public and policy discussions about the current state, necessity, and desirability of content moderation. We also hope to save others time by tracking down where policies about disinformation are housed. All the information within this document is accessible online and many parts of what we present here have been reported and condensed by the press. But, policies are not always located in the most intuitive places. For instance, rules against census and election interference on Facebook are found under "Coordinating Harm and Publicizing Crime" while Twitter has a stand-alone "Election Integrity Policy." The rules that are most applicable from Reddit are found under "Impersonation" in the platform's Content Rules, while they live in Snapchat's section on "Hate Speech and False Information" in its Community Standards. YouTube clarifies that census and voting misinformation is prohibited in its "Spam, deceptive practices, & scams policies." And, the COVID-19 pandemic has prompted many platforms to document changes in their policies through their company blogs and news centers rather than in the policy pages directly.



¹ We follow research distinguishing between 'disinformation' as content with the intent to deceive and cause harm, while 'misinformation' refers to content that is deceptive or false without that intent. We use 'misinformation' as the broader category here, and note when this distinction applies to a platform's policy. Freelon, Deen, and Chris Wells. "Disinformation as political communication." *Political Communication* 37, no. 2 (2020): 145-156.

Platform Approaches to Misinformation

Generally, it is perfectly acceptable for users to create and organically post factually incorrect information about politics across these platforms. However, there are four categories of content that social media platforms have differentially chosen to fact-check:

- Democratic processes
- Manipulated media
- Tragic events
- Health

These companies don't just carve out categories of content; they also carve out categories of accounts for varying standards and scrutiny. Broadly, these categories of users are:

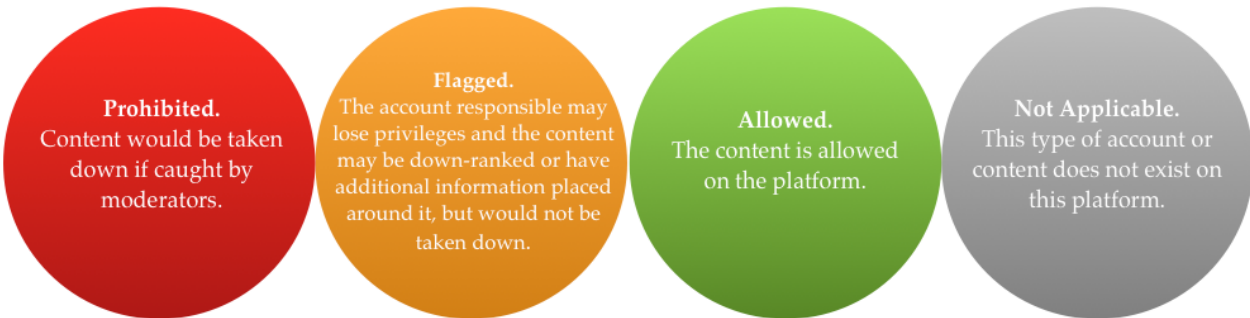
- Public figures
- Average users
- Monetized accounts/monetized content
- Advertisers/advertisements

This report takes six real examples of false information and shows how each platform would likely respond to it depending on who posted the content. Each platform varies in how categories of content and accounts are defined, but for the sake of interpretability we develop ideal types that fit all of their definitions and show how each platform would respond to a real case of mis- or disinformation.

Importantly, there is one last category of account that we do not repeat in our tables below: Facebook, Reddit, Snapchat, Twitter, and YouTube all prohibit using their platforms in coordinated or “inauthentic” ways. Facebook bans “inauthentic behavior” in its Community Standards while YouTube prohibits users to “cause or encourage any inaccurate measurements of genuine user engagement” in its terms of service. Twitter does not allow users to “artificially amplify or suppress information” in its platform manipulation policy. In Snapchat’s terms of service, users are only allowed one account per person and it is against the terms of service to give anyone else access. While Reddit has few rules, the platform does not allow vote manipulation, interfering with its normal use, and creating multiple accounts. This means that any misinformation posted by an “inauthentic” account is prohibited by the nature of the account.

Cases of misinformation often overlap with the multitude of other policies social media platforms have. We do not specifically address policies against bullying, harassment, hate speech, spam, or threats in this report unless they explicitly address false information, although such policies are often applicable to cases of misinformation.

Key



At the end of this document (Page 21), a numbered reference list contains all the platform policies associated with each cell in the tables.

*An asterisk means that the rules are dependent on third parties. In the case of Facebook, third-party fact-checkers are responsible for determining if a story is false, partly false, or true.

Types of Accounts

Platforms distinguish between accounts in important ways that have significant policy implications:

World leader

A 'world leader' meets the definition of a public figure or a politician on Facebook (both of which can be subject to different rules in the platforms' Community Standards and Fact-checking policies) as well as the definition of an elected or government official in Twitter's public-interest exception policy. On Facebook, politicians are exempt from fact-checks. On Twitter, tweets in the public-interest that would usually be taken down may instead be down-ranked in Twitter's algorithms and have contextual information placed around them.

Average Jane

Your average, non-public-figure user. Not a politician, government or elected official, owner of a monetized account, or advertiser.

Monetized account

Facebook, Snapchat, and YouTube give content creators the opportunity to share in the advertising revenue from the ads that appear next to their content. This includes publishers on Facebook, monetized video channels on YouTube, and Snapchat Discover partners.

Advertisements


























Advertisements are pieces of content that an advertiser pays the platform to show to more people than the content would reach organically. This includes sponsored posts on Facebook and promoted tweets on Twitter (typically not applicable in the given examples due to Twitter's ban on political advertising) as well as Reddit, Snapchat, and YouTube ads.

Types of Problematic Content

GENERAL PLATFORM FALSE INFORMATION POLICIES APPLIED TO POLITICAL ISSUES



Example One, False Information About a Social Issue. Source: *The New York Times*

					
WORLD LEADER					
AVERAGE JANE					
MONETIZED ACCOUNT					
AD					

Generally, social media platforms have no requirement that users tell the truth. This includes statements about candidates, elected officials, or social issues. The example above (this was not Christine Blasey-Ford) stands in for many potential cases, for example content saying there was looting at a Tea Party rally when no such actions occurred, or a post mis-attributing a quote from Benito Mussolini to Steven Mnuchin.

Such content would not run afoul of any platforms' rules when posted by a world leader or (on any platform except potentially Facebook) an everyday user. On YouTube, there may be "topical context" provided about the content, but it would still run without any sanctioning of the account that posted it or algorithmic downranking. This content would also be fine to monetize or run as a paid advertisement on YouTube. Topical context is not provided for ads on YouTube.

However, the post above would not be allowed as an ad on Reddit or Snapchat, and it would not be allowed on Snapchat's Discover section because of these platforms' broad rules that monetized content and ads must be "truthful" and not "false or misleading."

Facebook and Instagram are much more complicated than YouTube, Reddit, Snapchat, or Twitter. On these platforms, if the post had been fact-checked and deemed false by third-party fact-checkers, the content could not be run as an advertisement and any monetized account which posted it could lose privileges or be down-ranked in Facebook's newsfeed algorithm. Contextual information could be placed around it, such as the fact-check. However, these actions are dependent on a third-party fact-check occurring, which is not likely to happen to the vast majority of posts on the platform. One independent accounting of how many stories are checked by Facebook's third-party fact-checking partners found that a total of 302 stories were fact-checked in January of 2020. If no such fact-check occurs, the content would be allowed, without context and without policies for sanctioning or algorithmic downranking.

That said, all of this is contingent on the type of user. If an average user posted this content on Facebook or Instagram and that post was fact-checked, they would not be subject to losing privileges, but the content could include additional information around it alerting other users that it is false. World leaders are politicians and are thus exempt from fact-checking; if they posted the content above it would not be eligible for fact-checking and would be allowed. Because of this, they could also run it as an advertisement. For example, while the ads and posts of politicians can be demonstrably false, the ads of advocacy and civil society organizations or non-candidate political action committees can be fact-checked and removed, and their organically produced content downranked.





FALSE INFORMATION ABOUT DEMOCRATIC PROCESSES



Example Two, Election Misinformation.
Source: [Propublica/Twitter](#)



Example Three, Census Misinformation. Source: [Facebook](#)

					
WORLD LEADER	3		12	17, 19	23
AVERAGE JANE	3		12	17	23
MONETIZED ACCOUNT	3		12		20, 23
AD	3	8	12	14	21, 23

Social media platforms tend to draw hard lines when misinformation impacts elections and the census. Misinformation about how, where, and when to vote and who is on the ballot is prohibited on most platforms, as is misinformation designed to suppress the vote. Yet, platforms still have different scopes for election misinformation. For example, content from so-called ‘birthers’ who propagated the myth that Obama was not born in the US would likely have violated YouTube’s current candidate eligibility policies, but not necessarily other platforms current election integrity policies.

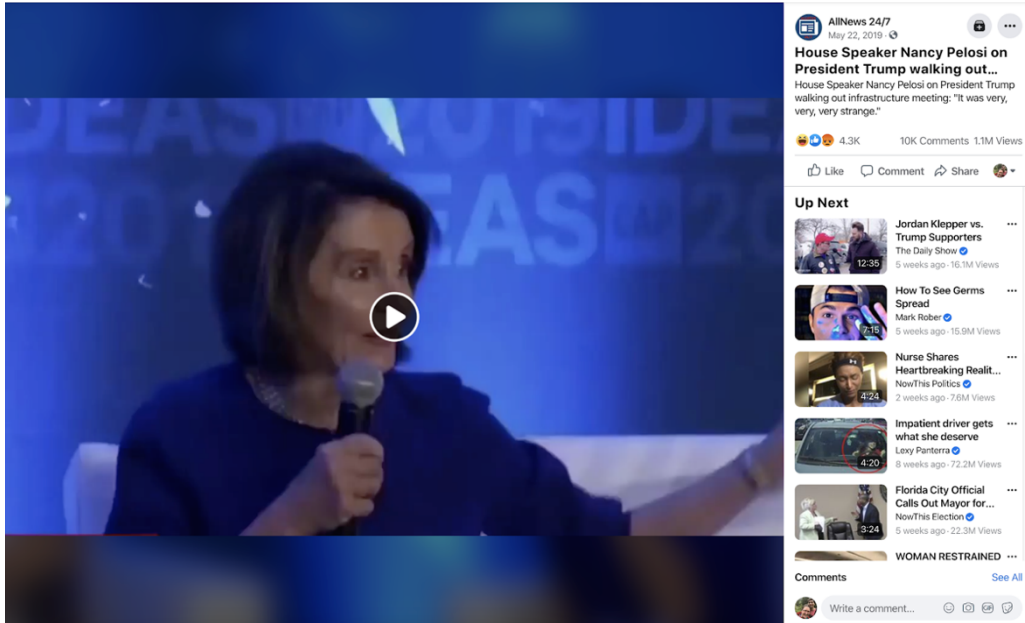
At the same time, false, inaccurate, or misleading information about the census is against most platforms’ policies, just like election misinformation. The census misinformation post above was a paid advertisement on Facebook run by the Trump campaign. It would be against Facebook and Instagram, YouTube, and Snapchat’s policies if run as an ad or organic content. After a journalist pointed out these ads on Facebook, they were taken down. On Twitter, such a post would be against policies but likely granted a public-interest exception if posted by a world leader (it would not be able to be run as an ad due to Twitter’s blanket ban on political ads). Under a public-interest exception, it would potentially be down-ranked to be viewed by less people and may have other restrictions placed on it.

The Community Standards on Facebook regarding election and census misinformation are not dependent on fact-checkers; the call is made instead by the platform itself. That’s why the Facebook column is entirely red instead of green or orange like the first example.






Reddit is the farthest from flat-out banning such content though exactly what the platform would take down is unclear due to the vagueness of its “impersonation” policy and the breadth with which the company has said it will be applied in conversations with the press. The policy is framed around “impersonation” and “manipulated content presented to mislead.” Census and election misinformation may only violate Reddit’s policy if it misrepresents who it is from or uses manipulated media to impersonate a source.

Facts about democratic processes should be straightforward. What the census is, how it works, and what is on its questionnaire should be easily verifiable, widely-agreed upon, and uncontested by major public figures and institutions. Norms in campaigning have traditionally prevented misinformation about democratic processes or the questioning of opponents' eligibility for office by major party leaders. As the examples above highlight, this may no longer be as much the case as social media platforms would hope.

MANIPULATED MEDIA



Example Four, Manipulated Media. Source: Facebook

					
WORLD LEADER	7	10		18, 19	23
AVERAGE JANE	4*	10		18	23
MONETIZED ACCOUNT	4, 6*		12		20, 23
AD	1*	10	11	14	21, 23

While mistruths are generally allowed across social media platforms, manipulated media is held to a higher standard. Manipulated media often have a clear, verifiable truth for social media platforms to compare against: the original media. So, while a user can say that Nancy Pelosi was drunk while giving a speech and no social media platform will verify the truth of that statement, they are much more likely to verify if a video posted is manipulated. If it is, that may be against some platforms' rules.

In the case of manipulated media, the definitions that social media platforms follow vary widely. Facebook, for instance, only applies its policy to manipulated media that is “the product of artificial intelligence or machine learning”; the company explicitly allows manipulated content that leaves out words or changes the order in which they were said provided these alterations were not done through AI or machine learning. This means that the video above is allowed, as are things such as photoshopped images. However, on Facebook, since the story that Nancy Pelosi was drunk has been fact-checked and deemed false, the video's distribution has been limited and it is not allowed in advertisements. Monetized accounts that share it could be subject to sanctions, including losing their monetized status.

In contrast, Reddit's broadly-worded impersonation policy against “manipulated content presented to mislead, or falsely attributed to an individual or entity” makes the Nancy Pelosi video against its rules. YouTube has also taken a broad definition of what counts as manipulated media with a strict requirement that it could cause harm: “Content that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm.” Youtube took down the “drunk” Nancy Pelosi video.

While Twitter may take content down if it is deceptively altered, shared in a deceptive way, and likely to cause serious harm, more often the content will only have context put around it, such as being labeled as manipulated media.


























YouTube prohibits false information about tragic events in its hate speech policy, which states that users cannot “deny that a well-documented, violent event took place.” This policy is about preventing harm, which to YouTube includes inciting hate and violence. Similarly, Snapchat’s Community Guidelines prohibit “deliberately spreading false information **that causes harm**, such as denying the existence of tragic events” (emphasis added). In both cases, the companies are banning this content to prevent harm.

While this type of disinformation is likely to overlap with Facebook, Twitter, and Reddit’s policies on hateful conduct, hate speech, bullying, or threats, none of these three platforms explicitly prohibit holocaust denial. In interviews on the subject, Mark Zuckerberg of Facebook insisted that users on his platform are allowed to make mistakes such as believing that the holocaust never occurred.

HEALTH INFORMATION



Example Six, Health Misinformation.

					
WORLD LEADER					
AVERAGE JANE					
MONETIZED ACCOUNT					
AD					

The example above is a screenshot from “Plandemic,” a 26-minute documentary-style video containing numerous false claims, some of which go directly against guidance from authoritative sources of public health information. However, the majority of the video consisted of false information about things such as Dr. Judy Mikovits’s career and conspiratorial statements about the government and scientific community. Facebook and YouTube removed the full video from their platforms, but Facebook has allowed shorter versions of it to remain as long as they do not contain the specific sentences that it deems have “the potential to contribute to real-world harm.” Those clips are eligible for fact-checking. Despite the video being clearly against Twitter’s rules and Twitter taking some actions (such as blocking hashtags from searches and trends and marking some related tweets as harmful in order to limit their spread), the video was still extremely easy to find on Twitter on Friday, May 8th.

During the coronavirus pandemic, social media platforms have all taken steps to promote accurate public health information. In addition to the changes designed to direct people towards good information (such as surfacing videos from public health officials first in results and creating shortcuts on home pages to facilitate access to information from public health institutions) and focusing content moderation efforts on coronavirus misinformation, Twitter, Facebook and YouTube have also expanded their rules against bad information. Twitter has broadened its definition of “harm” to include content “that goes directly against guidance from authoritative sources of global and local public health information” and has removed posts from the presidents of Brazil and Venezuela when they violated this policy. YouTube has also created a “COVID-19 Medical Misinformation Policy” that prohibits false information that “poses a serious risk of egregious harm” and contradicts the World Health Organization or local health authorities. Facebook has placed a panel on its Community Standards page about Covid-19 which specifies that it prohibits factually incorrect health content “that has the potential to contribute to real-world harm.” But, as the narrow scope of take downs of the Plandemic video show, the platform’s definition of “harm” can be a very limiting requirement.

Even when harm may be easily attributable, social media platforms still run into trouble. While Twitter has taken down tweets from two world-leaders containing coronavirus misinformation, it did not remove tweets from President Trump, including one that said “HYDROXYCHLOROQUINE & AZITHROMYCIN, taken together, have a real chance to be one of the biggest game changers in the history of medicine...” This statement of the efficacy of two drugs was not directly against the Center for Disease Control’s guidance only because the White House requested specific guidance language from the CDC.

All of these platforms’ efforts to enforce the truth not only rely on having the technological capabilities to do so, but also having widely agreed upon, uncontroversial facts from official sources and clear, real-world attributable harm to prevent. COVID-19 has highlighted the problems with all of these requirements. Like the examples from the census and election misinformation, in the United States misinformation is coming from the very institutions that platforms would expect to turn to for guidance on what is false. And the harms that are occurring are not always acute threats to someone’s life but instead the broad undermining of trust in health authorities.

Conclusion

Major social media platforms have negotiated their role in the spread of misinformation by generally prioritizing moderating content most likely to cause harm and only requiring factual accuracy in cases with clear institutional and scientific consensus, an original to compare against (in the case of manipulated media), or independent verification from fact-checkers for non-public figures (Facebook and Instagram). In this way, social media platforms can still avoid being arbiters of truth even while actively enforcing factual accuracy in important (if limited) categories of content. At the same time, it is clear that there is no consensus on 'harm' given only two platforms explicitly ban denial of the Holocaust

In addition, and not the explicit subject of this report, platforms have created a wide range of strategies beyond simply removing factually incorrect content. These strategies include down-ranking content and accounts in their algorithms, placing additional information around false information, promoting trustworthy sources, directing users who viewed or interacted with false information to good information, and requiring an additional click to view false information.

Unfortunately for Facebook, Reddit, Snapchat, Twitter, and YouTube, these strategies broadly rely on the existence of uncontroversial, established facts agreed upon by knowledge-producing and verifying institutions. Increasingly, however, such agreement is often not the case. As empirical studies show, [misinformation and its spread often comes from the top](#), including political leaders and media elites. This includes the typically trusted offices that the public, and social media platforms, should rely on for guidance. This leaves platforms in a tricky position: when world leaders, health officials, scientists, or government bodies disagree, whose truths should social media platforms enforce?

Policy References

Facebook and Instagram

This report draws primarily on Facebook’s Community Standards (which Instagram content is also governed by), Ads Policies, Fact-checking policies, and Content-monetization policies.

Advertising

Ads can be run on Facebook and Instagram as sponsored content within feeds and video rolls. In addition, ads run through Facebook’s Ad Manager run on over 10,000 websites and apps that partner with Facebook. Advertisements are not allowed to contain content that has been fact-checked as false.

Monetization

Video publishers and pages with over 10,000 followers can apply to monetization in which they receive a share of the advertising revenue from ads that run next to their content.

1. Advertising Policies – Misinformation:

https://www.facebook.com/policies/ads/prohibited_content/misinformation

2. Community Standards: <https://www.facebook.com/communitystandards/>

3. Community Standards - Coordinating Harm and Publicizing Crime:

https://www.facebook.com/communitystandards/coordinating_harm_publicizing_crime

4. Community Standards - False News:

https://www.facebook.com/communitystandards/false_news

5. Community Standards - Manipulated Media:

https://www.facebook.com/communitystandards/manipulated_media/

6. Content Monetization Policies:

<https://www.facebook.com/help/publisher/1348682518563619>

7. Fact-checking: <https://www.facebook.com/help/publisher/182222309230722>

Reddit

Reddit's content is governed at its base by Reddit's Content Rules, but the majority of moderation decisions are made at the subreddit level. Each sub has its own particular rules, often much more comprehensive and strict than Reddit's general Content Rules. Volunteer moderators of subreddits are primarily responsible for creating and enforcing these sub rules. Due to the wide variation in subreddits, this report focuses only on Reddit's Content Rules, which can be seen as the absolute minimum standards. Generally, Reddit heavily cites its "impersonation" policy in issues regarding misinformation, which prohibits "manipulated content presented to mislead" in addition to impersonation of people or other entities.

Advertising

Ads can be run on Reddit. Reddit's advertising policies require that ads be "truthful, non-deceptive, and defensible."

Content monetization

There is no content-monetization on Reddit; publishers do not get portions of revenue from advertisements run alongside their content.

8. Advertising Policy: <https://www.reddithelp.com/en/categories/advertising/ad-review/reddit-advertising-policy>

9. Content Policy: <https://www.redditinc.com/policies/content-policy>

10. Content Policy - Impersonation: <https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-impersonate-individual-or>

Snapshot

Snapchat has a unique structure that makes it much harder to discover content. There are no user-created public groups or forums to easily discover and join. Rather, content discovery is primarily limited to Snapchat's Discover platform which is vetted and curated by employees.

Advertising

Ads can be run on Snapchat. Snapchat's advertising policies include the broad requirement that advertisements be "truthful."

Content monetization

Publishers can make money off the advertisements run alongside their content on the Discover platform and they are held to much higher publication standards than on Facebook or YouTube.

11. Advertising Policy: <https://www.snap.com/en-US/ad-policies>
12. Community Guidelines: <https://www.snap.com/en-US/community-guidelines>
13. Snap News, including posts about Covid-19 efforts: <https://www.snap.com/en-US/news>

Twitter

Unlike Snapchat, there are no barriers to posting easily discoverable content on Twitter. There is no content monetization on Twitter.

Advertising

While ads can be run on Twitter, Twitter's political advertising ban prohibits many of the examples that are contained in this report regardless of the information in them.

Content monetization

There is no content monetization on Twitter.

14. Advertising Policy - Political Content: <https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html>

15. Advertising Policy - Inappropriate Content (Coronavirus):
<https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/inappropriate-content.html>

16. Coronavirus updates - expanding definition of "Harm":
https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#definition

17. Twitter Rules - Election Integrity: <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>

18. Twitter Rules - Synthetic and Manipulated Media: <https://help.twitter.com/en/rules-and-policies/manipulated-media>

19. Twitter rules - Public Interest Exception: <https://help.twitter.com/en/rules-and-policies/public-interest>

YouTube

Advertising

Ads are run through Google's advertising services, including Google Ads and Display and Video 360. These include pre-roll video ads as well as banners at the bottom of videos and YouTube search result ads.

Content monetization

Monetization on YouTube is common and is often the primary source of income for YouTube content creators.

20. Advertiser Friendly Content: <https://support.google.com/youtube/answer/6162278>

21. Advertising Policies: <https://support.google.com/adspolicy/answer/6008942>

22. Community Guidelines - Harmful or Dangerous Content Policy:

<https://support.google.com/youtube/answer/2801964>

23. Community Guidelines - Spam, Deceptive Practices, & Scams:

<https://support.google.com/youtube/answer/2801973>

24. Community Guidelines - Hate Speech Policy:

<https://support.google.com/youtube/answer/2801939>

25. Community Guidelines - Topical context:

<https://support.google.com/youtube/answer/9004474>

26. COVID-19 Medical Misinformation Policy:

<https://support.google.com/youtube/answer/9891785>