CELE

The New Normal?

Disinformation and Content Control on Social Media during COVID-19

April 2021

Facultad de Derecho
Centro de Estudios en Libertad
de Expresión y Acceso a la Información

UP
Universidad
de Palermo

The New Normal? Disinformation and Content Control on Social Media during COVID-19

Carlos Cortés and Luisa Fernanda Isaza
CELE

I. Presentation

This document addresses the measures that Facebook, Twitter, and YouTube implemented to address problematic content related to COVID-19 on their platforms. Known mainly as community guidelines, these rules are the basis for the moderation actions carried out by these services on user content. The main purpose of this paper is to understand the impact that COVID-19 had on community guidelines.

First, we describe what the status of the relevant regulations was before the pandemic. Second, we explain the changes these platforms made for organic content. Third, we address the public interest exception, which protects posts even if they violate community standards. Fourth, we describe some measures related to the content advertised on these services. Finally, we offer some conclusions.

This is not an exhaustive analysis on the subject, nor does it cover all the modifications that may have taken place in this matter. In fact, one of the problems identified throughout this research is the difficulty in understanding where they are, how they change, and how the community rules of social media are implemented. On the other hand, this document does not include topics such as deepfakes, manipulated multimedia content, or influence operations.

The study had a cutoff date of December 31, 2020, and made use of the monitoring of community guidelines carried out by CELE and Linterna Verde through the Letra Chica project.[1] Letra Chica tracks, explains and puts into context the changes to the community guidelines of Facebook, Twitter, and YouTube.[2] The measures covered by the text are also summarized in tables included as annexes.

This document was finished when a historical event was taking place as several social media platforms sanctioned the accounts of the then president of the United

[1] Linterna Verde is an interdisciplinary non-profit organization that answers questions about the digital public debate.

[2] For more information, see: https://letrachica.digital, last access: March 9, 2021.

States, Donald Trump. Specifically, Twitter suspended his account permanently while Facebook sent that same decision for review by the oversight board that became operational last year.[3] We should clarify, then, that this situation is not included in this text. However, the COVID-19 crisis coincided with the presidential campaign in the United States, and to that extent, some elements related to that debate are reflected in this study.

II. The world before the pandemic

Before the new coronavirus disease was declared a pandemic by the World Health Organization (WHO) in March 2020, disinformation was already a complex enough problem for Internet platforms. A month earlier, when the crisis was just beginning, the director of the WHO, Tedros Adhanom Ghebreyesus, warned: "This infodemic is hindering efforts to contain the outbreak, spreading unnecessary panic and confusion, and driving division."[4]

The elements pointed out by the WHO director could well be applied to what had been happening before this public health emergency: disinformation is an obstacle to containing the pandemic; it is used to spread panic and divides citizens.[5] Undoubtedly, the great point of tension prior to COVID-19 was the 2016 US presidential elections, where the Donald Trump campaign relied on a manipulation strategy whose impact on the result is still being studied.[6]

Of course, state pressure on the platforms increased. According to scholars from

[3] See, Clegg, Nick, "Referring Former President Trump's Suspension from Facebook to the Oversight Board," Facebook, January 21, 2021, retrieved from: https://about.fb.com/news/2021/01/referring-trump-suspension-to-oversight-board, last access: March 4, 2020.

[4] Adhanom Ghebreyesus, Tedros y NG, Alex, "Desinformación frente a medicina: hagamos frente a la 'infodemia'" [Disinformation against medicine: let's face the 'infodemic'], *El País*, February 18, 2020, retrieved from: https://elpais.com/sociedad/2020/02/18/actualidad/1582053544_191857.html, last access: March 4, 2020.

[5] On the measures taken by Facebook, Google and Twitter to combat disinformation before the pandemic, see: Cortés, Carlos and Isaza, Luisa, *Noticias falsas en Internet: la estrategia para combatir la desinformación* [Fake news on the Internet: The strategy to combat misinformation], CELE, December 2017, retrieved from: https://www.palermo.edu/cele/pdf/FakeNews.pdf, last access: March 4, 2020.

[6] Some studies with diverse conclusions on the influence of fake news on the elections: Gunther, Richard, Nisbet, Erik C. and Beck, Paul, "Trump May Owe his 2016 Victory to 'Fake News', New Study Suggests," The Conversation, February 15, 2018, retrieved from: https://theconversation.com/trump-may-owe-his-2016-victory-to-fake-news-new-study-suggests-91538, last access: March 4, 2020. Guess, Andrew M., Nyhan, Brendan and Reifler, Jason, "Exposure to Untrustworthy Websites in the 2016 US Election," Nature Human Behaviour, March 2, 2020, retrieved from: https://www.nature.com/articles/s41562-020-0833-x?proof=trueMay%252F, last access: March 4, 2020. Guess, Andrew, Nagler, Jonathan and Tucker, Joshua, "Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook," Science Advances, January 9, 2019, retrieved from: https://advances.sciencemag.org/content/5/1/eaau4586, last access: March 4, 2021

the University of Oxford, in the two years following the 2016 US elections, at least 43 countries around the world proposed or implemented regulations specifically designed to control operations of influence and disinformation.[7]

But criticism of social media sites — especially Facebook — for the disinformation that spreads through them is not limited to what happened in the US elections. In Burma (Myanmar), late and untimely access to the Internet (in 2014, 1% of the population used the Internet; in 2016, 20%),[8] coupled with an anti-Muslim propaganda strategy on social media, led to stigmatization, attacks, and violence against Rohingya Muslims.[9] In August 2018, following the murder of more than 25,000 Rohingya and the forced displacement of another 700,000, Facebook acknowledged that it had been slow to act on disinformation and anti-Muslim hatred in this country.[10] As a result, it belatedly suspended dozens of users and pages linked to the Burmese army — many with millions of followers — for violating the integrity and authenticity policies on the platform.[11]

In Latin America, the issues of misinformation and inauthentic activity on social media in electoral contexts have also been in the spotlight. The InternetLab research center analyzed the profiles of the Twitter followers of the presidential candidates in Brazil in 2018 to detect what percentage of them corresponded to potentially automated accounts: the candidate with the fewest number of bots had 13%, while the one with the most had reached 63%.[12] Furthermore, influence op-

---

[7] Bradshaw, Samantha, Neudert, Lisa-María and Howard, Philip N., *Government Responses to Malicious Use of Social Media*, Oxford Internet Institute, University of Oxford, 2019, retrieved from: https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/01/Nato-Report.pdf, last access: March 4, 2020. The study focused on the 100 countries with the highest number of Internet users in 2016, within which they identified the 43 countries mentioned. Within the regulation directed to social media, they found measures for the removal of content, transparency in political advertising, and the protection of personal data.

[8] Frenkel, Sheera, "This Is What Happens When Millions of People Suddenly Get the Internet," Buzz Feed News, November 20, 2016, retrieved from: https://www.buzzfeednews.com/article/sheerafrenkel/fake-news-spreads-trump-around-the-world, last access: March 4, 2021.

[9] *Ibid*. See also: Gowen, Annie and Bearak, Max, "Fake News on Facebook Fans the Flames of Hate against the Rohingya in Burma," *The Washington Post*, December 8, 2017, retrieved from https://www.washingtonpost.com/world/asia_pacific/fake-news-on-facebook-fans-the-flames-of-hate-against-the-rohingya-in-burma/2017/12/07/2c1fe830-ca1f-11e7-b506-8a10ed11ecf5_story.html, last access: March 4, 2021.

[10] Facebook, "Removing Myanmar Military Officials from Facebook," August 28, 2018, retrieved from: https://about.fb.com/news/2018/08/removing-myanmar-officials, last access: March 4, 2020.

[11] In August 2020, Facebook announced how it is preparing to combat hate speech and disinformation ahead of the Burma elections in November 2020, the second democratic elections in recent Burma history. Frankel, Rafael, "How Facebook Is Preparing for Myanmar's 2020 Election," Facebook, August 31, 2020, retrieved from: https://about.fb.com/news/2020/08/preparing-for-myanmars-2020-election, last access: March 4, 2021.

[12] Lago, Lucas and Massaro, Heloisa, "Bot or not: Who Are the Followers of our Candidates for President?" InternetLab, retrieved from: https://www.internetlab.org.br/en/information-politics/bot-or-not-who-are-the-followers-of-our-candidates-for-president, last access: March 4, 2021.

erations in social media are common in countries such as Mexico, Colombia, and El Salvador, among others, where "call center" accounts identify tags and trends and launch coordinated attacks on critics, journalists, and political leaders.

Despite the complaints and pressures from different sectors, the platforms had been taking a rather passive role in the face of content causing disinformation, by focusing their intervention on the authenticity of the accounts and the visibility of the publications. The leaders of these companies argued, with good reason, that these platforms should avoid becoming judges of the public debate.

Twitter was the one that resisted that role the most: "What we could do is help provide more context, either by showing all the different perspectives (…). But I think it would be dangerous for a company like ours to be arbiters of the truth," said Twitter founder and CEO Jack Dorsey in August 2018.[13] Along the same lines, Mark Zuckerberg, founder and CEO of Facebook, argued — in a well-known speech given at Georgetown University in October 2019 — that although he was concerned about the erosion of truth, he did not believe that people would like to "live in a world where you can only post things that tech companies judge to be 100% true."[14]

The center of the debate for YouTube has been different: besides the misinformation of content, there is the platform's recommendation system, which creates incentives for similar publications to be produced and consumed, with which the user experience ends up dominated by a rabbit hole effect.[15] In December 2019, Susan Wojcicki, CEO of YouTube, said on this point that a strong intervention of the platform could harm the availability of relevant content for users: "If we were held liable for every single piece of content that we recommended, we would have to review it. That would mean there would be a much smaller set of information that people would be finding. Much, much smaller."[16]

---

[13] "*I think what we could do is help provide more context, whether it be showing all the different perspectives (…). We have not figured this out, but I do think it would be dangerous for a company like ours… to be arbiters of the truth*". . Interview with CNN, August 2018, retrieved from: https://youtu.be/Cm_lmWWKDug?t=503, last access: March 4, 2020.

[14] "*While I certainly worry about an erosion of truth, I don't think most people want to live in a world where you can only post things that tech companies judge to be 100 percent true*". Georgetown University speech, October 2019, retrieved from: https://youtu.be/2MTpd7YOnyU?t=1802, last access: March 4, 2021.

[15] Fox, Chris, "YouTube: 'We Don't Take You down the Rabbit Hole'", BBC News, July 19, 2019, retrieved from: https://www.bbc.com/news/technology-49038155, last access: March 4, 2021.

[16] "*If we were held liable for every single piece of content that we recommended, we would have to review it. That would mean there'd be a much smaller set of information that people would be finding. Much, much smaller*". "60 Minutes" Interview. Loizos, Connie, "In '60 Minutes' Appearance, YouTube's CEO Offers a Master Class in Moral Equivalency," Tech Crunch, December 1, 2019, retrieved from: https://techcrunch.com/2019/12/01/in-60-minutes-appearance-youtubes-ceo-offers-a-master-class-in-moral-equivalency, last access: March 4, 2021.

The position of the leaders of these companies is explained from a commercial and political perspective: arbitrating content was already a costly and time-consuming task, and trying to evaluate the veracity of information would end up confronting social media platforms with political parties, governments, and civil society. However, there were also substantive grounds on freedom of expression, its due process, and transparency.

In any case, in the pre-pandemic world, social media approached disinformation delicately. In broad terms, its principle had been that of minimal intervention in the content, either because there were no rules on the subject or because of their inconsistent application. In the words of journalist Casey Newton, "Every tech platform has two policies about what they will allow: the policy that's written, and the policy that's enforced. Ideally there would be no gap between these, but in practice it almost can't be helped."[17]

Faced with the Coronavirus emergency, social media changed its focus on the fly, designing and implementing measures that until recently seemed impossible. To do this, they used two types of rules: those that focus on identifying inauthentic actions on the platform and those that aim to judge the content. Based on their community rules, platforms have content moderation processes that rely heavily on human analysis. These are real armies of people trained to make complex decisions on a massive scale, without sufficient context, and under a lot of emotional pressure. And if this task was difficult when the focus was theoretically on simpler issues — such as identity theft, disclosure of private information, or threats — the misinformation surrounding the pandemic simply brought out how insurmountable the problem was.

Automation does not offer, at least for now, a solution either. Although algorithms can detect suspicious behavior, spam, and certain types of content — especially regarding videos and photos — they do not know how to solve dilemmas inherent to the context of an expression, much less decide on the veracity of information. The underlying risk is to end up with a high percentage of "false positives," with the consequent negative effects in terms of inhibition and censorship.

In essence, COVID-19 appeared as social media was putting out fires, prioritizing the most relevant markets and the most damaging news scandals. This mode of

[17] Newton, Casey, "Getting Rid of QAnon Won't Be as Easy as Twitter Might Think," The Verge, July 23, 2020, retrieved from: https://www.theverge.com/interface/2020/7/23/21334255/twitter-qanon-ban-facebook-policy-enforcement-political-candidates, last access: March 4, 2021.

damage control has been characterized by inconsistent rules and processes with slow transparency progresses. Next, we will see what the rules of Facebook, Twitter, and YouTube were for dealing with disinformation before the pandemic hit.

1. Facebook

In Facebook, in principle, the publication of fake news does not violate its community rules.[18] Therefore, it does not remove this content, but rather reduces its distribution.[19] And to do so, the company relies primarily on advanced third-party verification.

In a process known as fact-checking, civil society organizations around the world evaluate thousands of publications, mainly from the media.[20] With this input, Facebook implements some action that affects the visibility and distribution of the evaluated content: it alerts the person who sees it or who is going to share it; it "penalizes" by partially or totally reducing its visibility in the news section, and can even penalize accounts that create or share it repeatedly.[21] These measures, however, are not consistent across the world.

Some other rules and announcements are relevant:

- Facebook has a section of rules on "coordinating harm and publicizing crime."[22] It bans the spreading of content that shows, confesses, or encourages acts of physical harm to human beings, false calls to emergency services, or participation in high-risk viral challenges.[23]
- In the chapter on "Regulated goods," Facebook prohibits the publication of content that seeks to buy, sell, trade, donate, gift, or solicit drugs.

---

[18] "*False news does not violate our Community Standards*." Lyons, Tessa, "Hard Questions: What's Facebook's Strategy for Stopping False News?" Facebook, May 23, 2018, retrieved from: https://about.fb.com/news/2018/05/hard-questions-false-news, last access: March 4, 2021.

[19] Facebook, "21. Noticias falsas" [21. False new], retrieved from: https://www.facebook.com/communitystandards/false_news, last access: March 4, 2021.

[20] More information on the fact-checking program and Facebook's strategy to reduce fake news: Facebook, "Verificación de datos en Facebook" [Fact-checking in Facebook], retrieved from: https://www.facebook.com/business/help/2593586717571940, last access: March 4, 2021, and Lyons, "Hard Questions: What's Facebook's Strategy for Stopping False News?" *op. cit*.

[21] For more detailed information on the measures taken by Facebook to control disinformation immediately after the 2016 US elections, see: Cortés and Isaza, *Noticias falsas en Internet: la estrategia para combatir la desinformación*, [Fake news on the Internet: the strategy to battle misinformation] *op. cit*.

[22] Facebook, "2. Personas y organizaciones peligrosas" [Dangerous people and organizations], retrieved from: https://www.facebook.com/communitystandards/dangerous_individuals_organizations, last access: March 4, 2021.

[23] With this policy, Facebook prohibits, for example, the incitement to make false calls to the emergency services (an activity known as swatting) or to participate in high-risk viral challenges.

- In a blog post — that is, outside the community standards — Facebook reported that disinformation about health that may contribute to imminent physical harm was being deleted since 2018.[24] In this same post, it later clarified that since January 2020 it has deleted false publications about COVID-19. We will return to this point later.

Beyond these elements, Facebook has aimed to control misinformation by focusing on identifying "information or influence operations," i.e., coordinated actions that use automated and human accounts to amplify content, intimidate other users, and capture conversations and trends. We should remember that this discussion was framed in serious accusations of Russian interference in the 2016 United States elections.

In October 2019, Mark Zuckerberg claimed that this was "a much better solution than the ever-expanding definition of what constitutes harmful speech."[25] Accordingly, for Zuckerberg, the real problem with the Russian publications that sought to interfere in the elections was not their content, but the fact that they were made in a coordinated manner by fake accounts.

By early 2020, Facebook already had rules on misrepresentation and inauthentic behavior.[26] First — unlike Twitter — Facebook has a policy of real names, which, in the first place, requires that a person use their legal name when they register an account; second, the platform does not allow its users to publish, interact with content or create accounts with a high frequency; third, there are anti-spam rules: users cannot require or trick others to interact with certain content. Finally, Facebook prohibits engaging in inauthentic behavior, individual or coordinated, to mislead people regarding the popularity or origin of content.[27]

---

[24] Clegg, Nick, "Combating Covid-19 Misinformation across our Apps," Facebook, May 25, 2020, retrieved from: https://about.fb.com/news/2020/03/combating-covid-19-misinformation, last access: March 5, 2020. According to Facebook, since 2018 it has deleted, among others, content on measles in Samoa and the polio vaccine in Pakistan.

[25] Romm, Tony, "Zuckerberg: Standing for Voice and Free Expression," *The Washington Post*, October 17, 2019, retrieved from: https://www.washingtonpost.com/technology/2019/10/17/zuckerberg-standing-voice-free-expression, last access: March 5, 2020. Sarah C. Haan presents a position quite critical of this strategy in: Haan, Sarah C., "The Authenticity Trap. Mark Zuckerberg Thinks Facebook's Problems Can Be Fixed with 'Authentic' Speech. He's so Wrong," Slate, October 21, 2019, retrieved from: https://slate.com/technology/2019/10/mark-zuckerberg-facebook-georgetown-speech-authentic.html, last access: March 5, 2021.

[26] Facebook, "17. Integridad de la cuenta y autenticidad de identidad" [Account Integrity and Authentic Identity], retrieved from: https://www.facebook.com/communitystandards/integrity_authenticity, last access: March 5, 2020.

[27] Facebook, "20. Comportamiento no auténtico" [Inauthentic Behavior], retrieved from: https://www.facebook.com/communitystandards/inauthentic_behavior, last access: March 20, 2021.

2. Twitter

Following the line raised by Jack Dorsey according to which Twitter should not become an arbiter of the truth, the platform wanted to confront misinformation by focusing on the activity of the accounts. In other words, it focused on the actors rather than the content.

During the 2016 presidential campaign, the use of bots on Twitter did not occur in isolation. In June 2017, when the responsibility of social media in the Russian interference of the previous year in the United States was discussed, Twitter vindicated the open nature of its platform and announced that it would concentrate its efforts on avoiding automation for manipulation purposes: "We strictly prohibit the use of bots and other networks of manipulation to undermine the core functionality of our service," stated the company on its official blog.[28] In this context, in 2018 Twitter's objective was to detect these operations to avoid disinformation, attacks, and spam.[29]

However, before 2020 Twitter did begin to include some prohibitions related to disinformation, specifically in the context of electoral processes. In 2019, Twitter established a policy on "Election Integrity" prohibiting the posting of misleading information about how to participate, voter suppression and intimidation content, and false or misleading information about political affiliation.[30] That same year, on the occasion of the elections in the European Union, the platform introduced the possibility for users to report tweets that violate this policy. As has been a constant among these companies, the new feature had a geographic focus and did not imply a change in community standards. In 2020, with the US census and the presidential election campaign on the horizon, Twitter expanded the rules and renamed them "civic integrity policy."

[28] Crowell, Colin, "Our Approach to Bots and Misinformation," Twitter, June 14, 2017, retrieved from: https://blog.twitter.com/official/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html, last access: March 5, 2021.

[29] Twitter Public Policy, July 30, 2020, retrieved from: https://about.twitter.com/en_us/advocacy/elections-integrity.html#us-elections, last access: March 5, 2020. Currently, Twitter addresses this issue in reports about tampering with the platform. The latest available is from the second half of 2019: Twitter, "Platform Manipulation," 2019, retrieved from: https://transparency.twitter.com/en/reports/platform-manipulation.html#2019-jul-dec, last access: March 5, 2021.

[30] European Commission, "Fourth Intermediate Results of the EU Code of Practice against Disinformation," May 17, 2019, retrieved from: https://ec.europa.eu/digital-single-market/en/news/fourth-intermediate-results-eu-code-practice-against-disinformation, last access: March 5, 2020. See Twitter's report from April 2019, "Platform Manipulation," *op. cit*.

3. YouTube

YouTube's strategy to control disinformation is based on three principles: i) the preservation of content on the platform unless it violates its community guidelines; ii) the possibility of monetizing publications is a privilege, and iii) the videos must meet exacting standards for the platform to recommend them.[31] For the first of these principles, regarding the control of disinformation about the coronavirus disease, several community regulations would be relevant:

- In its policy on harmful or dangerous content, YouTube prohibits publications that promote, recommend, or assert that the use of harmful substances or treatments may have health benefits.[32]
- In its policy on deceptive practices, the platform prohibits manipulated or modified content that seeks to deceive the user and that may involve serious risk of blatant harm.[33]
- YouTube has a broader set of rules to ensure authentic behavior. Throughout different policies,[34] YouTube prohibits various types of behavior that deceptively seek to redirect users to other sites or artificially increase engagement metrics (views, comments, "likes"), as well as creating playlists with misleading titles or descriptions that make users believe that they will watch different videos than those in the list.

To ensure higher quality content and combat spam and other deceptive actions, for its second principle, YouTube postulates that only quality videos can be monetized.[35] In other words, creators seeking to monetize their videos must not only adhere to general community standards but also must adhere to stricter rules on

[31] Google, "How Google Fights Disinformation," February 2019, retrieved from: https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf?hl=en, last access: March 5, 2021.

[32] Google, "Política de contenido perjudicial o peligroso" [Harmful or Dangerous Content Policy], retrieved from: https://support.google.com/youtube/answer/2801964?hl=es-419&ref_topic=9282436, last access: March 5, 2021.

[33] Google, "Políticas sobre spam, prácticas engañosas y estafas" [Spam, Deceptive Practices and Scams Policies], retrieved from: https://support.google.com/youtube/answer/2801973?hl=es-419&ref_topic=9282365, last access: March 5, 2021.

[34] *Ibid*. Google, "Política de participación falsa" [Misrepresentation Policy], retrieved from: https://support.google.com/youtube/answer/3399767?hl=es-419, last access: March 5, 2020. Google, "Política sobre las listas de reproducción" [Playlists Policy], retrieved from: https://support.google.com/youtube/answer/9713446?hl=es-419&ref_topic=9282365, last access: March 5, 2021.

[35] To get money from the advertising the platform displays on its videos, creators must be part of YouTube's partner program. To belong to the program, at least a thousand subscribers and four thousand hours of video are required. Google, "Descripción general y elegibilidad del programa de socios de YouTube" [YouTube Partner Program Overview and Eligibility,], retrieved from: https://support.google.com/youtube/answer/72851?hl=es-419, last access: March 5, 2021.

advertiser-friendly content. A final YouTube strategy to improve the quality of the videos that users consume is to enrich the recommendation system. Based on the history of the users and other elements of the algorithm, the platform makes personalized suggestions on the home page, in the search results, and in the "Up Next" section that appears when a video is ending.[36]

III. More pressure, more measures

COVID-19 became the perfect storm for platforms. The increasing pressure they had been facing to remove problematic content became a public health issue. Disinformation surrounding the pandemic spread like the virus — at the hands of political leaders and influencers — community social media guidelines were insufficient and inconsistent, and those responsible for enforcing them had to confine themselves to their homes. Amid the confusion, platforms quickly began announcing new rules and measures to deal with misinformation. Regarding community rules, the actions reported during the pandemic have focused more on the rules on the content of the publications, than on rules about inauthentic activities.[37] Below is a description of the actions that the platforms are taking.[38] In a later section, there is an account of the new advertising rules that companies have established.

1. Facebook

In January 2020, when the coronavirus disease was not yet a global pandemic, Facebook was the first social media platform to make announcements about controlling disinformation. Facebook explained that at that time the platform used rules that were already in place. In essence, the intervention revolved around labeling, filtering, and content removal:

[36] Initially, the recommendation system suggested the content that achieved the most clicks by users. Consequently, YouTube changed the approach: to identify the content that users like the most, it looks at the time they spend watching the video and if they watch until the end. On the other hand, YouTube also seeks to detect and give prevalence to reliable sources, for example, taking the number of links that point to content as an indicator of authority. Finally, YouTube also conducts surveys to study whether users are satisfied with the recommendations. Google, "How Google Fights Disinformation," op. cit.

[37] In blog posts in January and March 2020, Twitter maintained that at the time it was not seeing significant coordinated platform manipulation efforts around the coronavirus disease. Twitter, "Our Zero-tolerance Approach to Platform Manipulation," March 4, 2020, retrieved from: https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#zero-tolerance, last access: March 11, 2021.

[38] For a detailed review of the changes see: CELE, Letra Chica, "Cambios por covid-19" [Changes due to COVID-19], Update from February 12, 2021, retrieved from: https://letrachica.digital/wiki/cambios-covid, last access: March 5, 2021.

- Facebook continues working with external fact-checkers.[39] Based on the verifiers' diagnosis, Facebook labels false information and limits its dissemination on the platform. Once a piece of content is flagged, Facebook activates proactive detection methods for possible duplicates.[40]
- Following a practice implemented before the pandemic, the platform alerts people who have shared or are trying to share content marked as false.
- Facebook promised to remove false content and conspiracy theories that could cause harm to people who believe them and which contradicted health authorities. The platform explained that the focus is on the content that discourages proper treatment or the implementation of prevention measures.

In its blog, Facebook reported that, following the guidelines of the WHO and other health authorities, since January it had deleted false information about cures, treatments, availability of essential services, the location and severity of the outbreak, etc.[41] In that post, the platform explained that this is not a new practice: since 2018 it has eliminated misinformation that may cause imminent physical harm, such as false information about measles in Samoa or rumors about the polio vaccine in Pakistan.

According to the company, these removal actions are done as an extension of an existing rule about harm-inducing content. And although that norm could be interpreted in that way, this is not directly inferred nor is it necessarily obvious.[42]

---

[39] "Nuestra red global de verificadores de datos externos continúa su trabajo revisando el contenido y desacreditando las afirmaciones falsas que se están extendiendo relacionadas con el coronavirus" [Our global network of external fact-checkers continues their work by reviewing content and debunking the spread of false claims related to the Coronavirus disease], the company explained on the blog on the subject. Jin, Kang-Xing, "Keeping People Safe and Informed about the Coronavirus," Facebook, December 18, 2020, "Limiting Misinformation and Harmful Content," January 30, 2020, retrieved from: https://about.fb.com/news/2020/08/coronavirus, last access: March 5, 2021.

[40] Rosen, Guy, "An Update on our Work to Keep People Informed and Limit Misinformation about Covid-19," Facebook, April 16, 2020, retrieved from: https://about.fb.com/news/2020/04/covid-19-misinfo-update, last access: March 5, 2021.

[41] Clegg, Nick, "Combating Covid-19 Misinformation across our Apps," *op. cit*.

[42] "Política sobre organización de actos para infringir daños y publicidad de la delincuencia: daños a personas. Mostrar, confesar o fomentar los siguientes actos cometidos por ti o personas relacionadas contigo: actos de daños físicos hacia seres humanos, incluidos actos de violencia doméstica, excepto cuando se comparte en un contexto de redención o defensa de uno mismo u otra persona. Declarar la intención de realizar, incitar, representar, apoyar o defender, o mostrar, confesar o fomentar los siguientes actos cometidos por ti o personas relacionadas contigo: llamadas falsas a los servicios de emergencia (*swatting*). Mostrar, fomentar, defender o incitar: la participación en desafíos virales de alto riesgo" [Policy on Coordinating Harm and Publicizing Crime: harm to people. Show, confess or encourage the following acts committed by you or people related to you: acts of physical harm towards human beings, including acts of domestic violence, except when shared in a context of redemption or self-defense or of another person. Declare the intention to make, incite, represent, support or defend, or show, confess or encourage the following acts committed by you or people related to you: false calls to emergency services (swatting). Show, encourage, defend or incite: participation in high-risk viral challenges]. Facebook, "3. Organización de actos

A month after the first statement, the measures began to include other aspects, with problems with COVID-19 content on the horizon. For example, Facebook announced that it would disable the option to search for virus-related augmented reality effects on Instagram; excluded from the recommendations content or organic accounts related to COVID-19; and, as if it were a notification of exposure to the virus, Facebook reported that it would alert people who had interacted (with "likes", reactions or comments) with content that had been discredited — including recommendations for reliable information. In practice, this approach allowed Facebook to create a series of informal rules on the fly to deal with disinformation during the pandemic, which it also extended to Instagram.[43]

Towards the end of the year, in light of the approval of COVID-19 vaccines, Facebook announced that it would remove false claims about COVID-19 vaccines that were disproved by public health experts.

2. Twitter

Twitter did a 180-degree turn: it went from being the platform that least intervened in user content to the most active. In March 2020, it warned that it would expand its definition of harm to include content that goes directly against the instructions of authorized sources in global and local public health.[44] Twitter then prohibited tweets that invite these behaviors or have the following content:

- Denying the recommendations of the authorities with the intention that people act against them.
- Encouraging breaking social distancing.
- Recommending ineffective treatments, even if they are not harmful or if they are shared humorously.

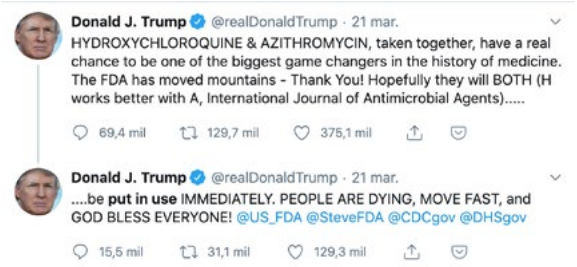
para infringir daños y promoción de la delincuencia" [Coordinating Harm and Publicizing Crime], retrieved from: https://www.facebook.com/communitystandards/coordinating_harm_publicizing_crime, last access: March 20, 2021

[43] Facebook has announced similar measures for Instagram, one of its products: i) blocking or restricting the use of hashtags to spread disinformation; ii) accounts or content related to COVID-19 are removed from the recommended section, unless they come from trusted health organizations; iii) the option to search for augmented reality effects related to COVID-19 is disabled, unless they have been developed in partnership with recognized health organizations. Jin, Kang-Xing, "Keeping People Safe and Informed about the Coronavirus," Facebook, December 18, 2020, retrieved from: https://about.fb.com/news/2020/10/coronavirus, last access: March 5, 2020.

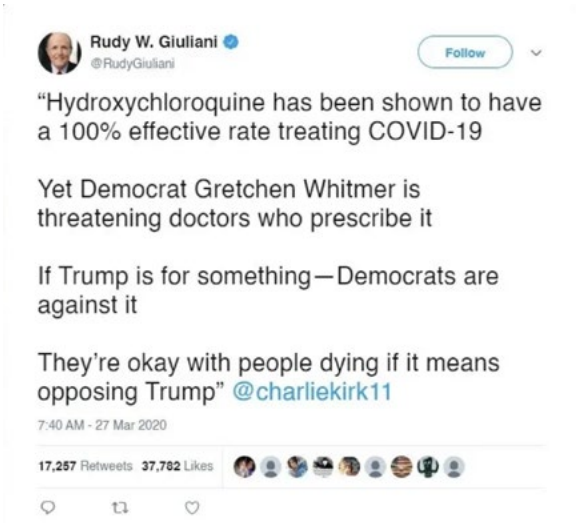
[44] Gadde, Vijaya and Derella, Matt, "Nueva información sobre nuestra estrategia continua sobre el covid-19" [New information on our ongoing strategy on COVID-19], Twitter, March 17, 2020, retrieved from: https://blog.twitter.com/es_la/topics/company/2020/nueva-informacion-sobre-nuestra-estrategia-continua-sobre-el-covid-19.html, last access: March 5, 2021.

- Recommending harmful treatments.
- Denying scientific data on the transmission of the disease.
- Making statements that incite action, cause panic, discomfort, or disorder on a large scale.
- Including false or misleading information about the diagnosis.
- Making claims that specific groups or nationalities are not susceptible or are more susceptible to the virus.

The implementation of these measures was not consistent. While a couple of tweets from the then president of the United States, Donald Trump, did not warrant any response from Twitter — a decision that, as we will see later, would be explained by reasons of public interest — the platform did act on a tweet from the president of Brazil, Jair Bolsonaro, and even temporarily suspended the account of Trump's lawyer, Rudy Giuliani, for similar reasons.[45]



Thread published by President Trump on March 21, 2020 where he promotes unproven medical treatments against the Coronavirus disease.



This tweet promoting the use of hydroxychloroquine, published on March 27, 2020, was deleted by Twitter and earned Giuliani the temporary suspension of his account.

[45] President Bolsonaro published on March 29, 2020, two videos that were deleted by Twitter, in which he was seen visiting public places in Brasilia, where he encouraged people not to self-isolate and spoke in favor of the use of hydroxychloroquine to treat the virus. The removal came days after the announcement of the new types of prohibited content. Darlington, Shasta, "Twitter elimina publicaciones sobre coronavirus del presidente de Brasil, Jair Bolsonaro" [Twitter removes posts about the coronavirus disease from the president of Brazil, Jair Bolsonaro], CNN, March 30, 2020, retrieved from: https://cnnespanol.cnn.com/2020/03/30/twitter-elimina-publicaciones-sobre-coronavirus-del-presidente-de-brasil-jair-bolsonaro, last access: March 5, 2021.

In May 2020, Twitter updated its approach to misleading information related to COVID-19 and explained that it takes action on content based on three categories:[46]

- Misleading information: statements or claims that have been confirmed as false or misleading by experts in the field, such as public health authorities.
- Controversial statements: statements or claims in which the accuracy, veracity, or credibility of the statement is questioned or unknown.
- Unverified claims: information that has not been confirmed at the time it is shared.[47]

In each type of content, the propensity to harm can be moderate or severe. According to Twitter, the removal of posts is done in case of misleading information with a propensity for severe harm. For the other cases, they use labels and filters:[48]

Misleading information on Twitter		
Categories \ propensity to harm	**Moderate**	**Severe**
Misleading information	Filter	Removal
Controversial claims	Labels	Filter
Unverified claims	No action	No action in principle. Label as necessary.

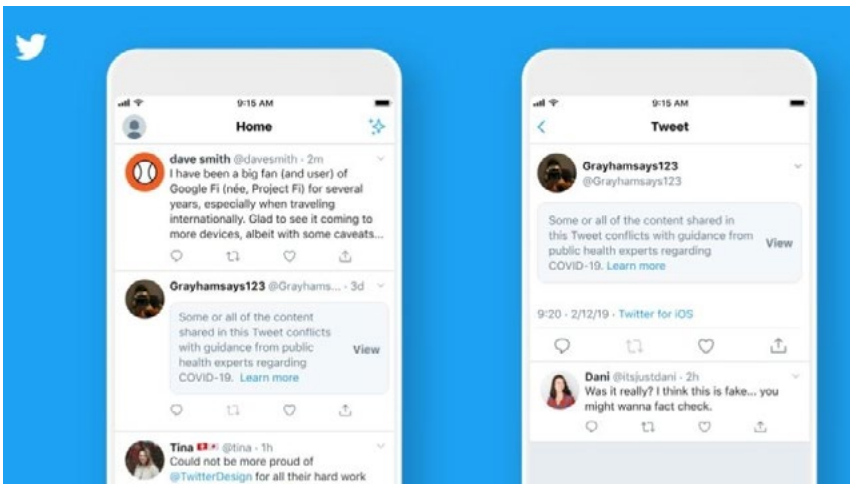
Filters or notices hide the questioned tweet to notify the user that the content differs from the guidance of public health experts. Following the platform's explanation, these apply in cases of misleading information with a moderate propensity to harm or controversial claims with a severe propensity to harm.

Labels, for their part, appear as phrases below the tweet, accompanied by an exclamation mark, referring to reliable information. Labels can be used in cases of controversial claims with a moderate propensity to harm, and in some cases of unverified claims with a severe propensity to harm.

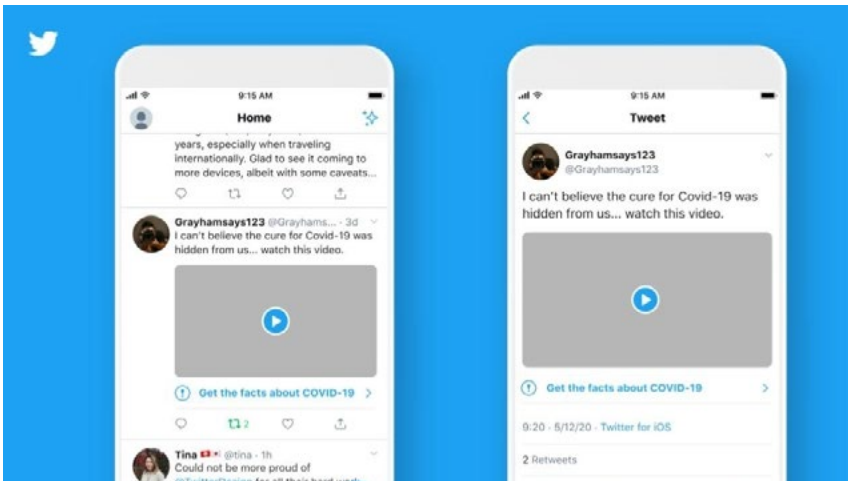
[46] Roth, Yoel and Pickles, Nick, "Actualizamos nuestro enfoque sobre información engañosa" [We updated our approach to misleading information], Twitter, May 11, 2020, retrieved from: https://blog.twitter.com/es_la/topics/product/2020/actualizamos-nuestro-enfoque-sobre-informacion-enganosa.html, last access: March 5, 2021.

[47] *Ibid*.

[48] For filters, Twitter uses the word "notice." Yoel and Pickles, "Actualizamos nuestro enfoque sobre información engañosa" [We updated our approach to misleading information], *op. cit*.



Example of filter or "notice."



Example of label.

Like Facebook, Twitter announced these rules in the company's blog posts. However, to date they have not been incorporated into their community standards, nor is it clear if they will be implemented in the same way in other similar cases.

3. YouTube

YouTube, which already had regulations against content that recommend the use of dangerous substances, created a new "policy on medical misinformation related to COVID-19."[49] These rules were not announced but were directly incor-

[49] Google, "Política sobre información médica errónea relacionada con el covid-19" [COVID-19 medical misinformation policy], retrieved from: https://support.google.com/youtube/answer/9891785, last access: March 5, 2021.

porated into their community guidelines. Content about COVID-19 that implies a "serious risk of flagrant harm" is prohibited. Under these rules, YouTube does not allow content that discloses erroneous medical information that contradicts the guidelines of the WHO or local health authorities regarding the treatment, prevention, diagnosis, or transmission of the virus. For example, it is prohibited to state that there have been no deaths from COVID-19 or that there is a guaranteed cure against the virus; hold that certain people are immune to the virus because of their race or nationality; discouraging users from consulting a medical professional if they become ill or encouraging them to turn to home remedies instead of seeking medical attention; assert that the origin of the virus is in the 5G networks; declare that social distancing is not an effective measure to reduce the spread of the virus; and state that the vaccine against the virus will cause death."

Violation of the policy results in the removal of content. In addition, YouTube applies a strikes system: with the first violation, the user receives a warning. After the first time, YouTube adds a strike. Three strikes result in the permanent removal of the channel. In this policy, YouTube also included the exception for educational, documentary, scientific, or artistic purposes. According to the policy, content that is in principle prohibited may be allowed if the content includes context "that gives equal or greater weight to countervailing views from local health authorities or to medical or scientific consensus." Exceptions can also be made if the purpose of the content is to condemn or dispute misinformation, by providing justifying context. The next section explains more about the exceptions that platforms apply to their own rules.

IV. Public Interest and Similar Exceptions

International human rights standards defend in particular expressions related to the public interest, which implies that they are granted a higher threshold of protection. The Inter-American Court of Human Rights links public interest to matters related to the functioning of democracy and the state, as well as to public management and the exercise of rights.[50] Similarly, social media platforms pro-

[50] "En cuanto al carácter de interés público, en su jurisprudencia la Corte ha reafirmado la protección a la libertad de expresión respecto de las opiniones o informaciones sobre asuntos en los cuales la sociedad tiene un legítimo interés de mantenerse informada, de conocer lo que incide sobre el funcionamiento del Estado, o afecta derechos o intereses generales o le acarrea consecuencias importantes" [Regarding the nature of public interest, the Court has confirmed the protection of freedom of expression regarding opinions or information on matters in which society has a legitimate interest in being informed, in knowing what affects the functioning of the state, or affects rights or general interests or entails important consequences]. I/A Court

tect certain types of expressions as they are considered to be of public interest. Therefore, it is the exception that the platforms apply to content that may be in violation of their rules but which they decide not to remove because they consider that it is in the interest of the public.

In general, the criteria for the use of the exception focus on the subject that produces the content. If a person says that children are immune to the Coronavirus disease, this post can be removed for violating Twitter's rules. But the situation may change if the person who expresses it is Donald Trump since users have a high interest in learning about the views of the president of the United States.

It is a difficult principle to apply, which in the case of social media becomes even more confusing. As we have already seen, the platforms have not granted the President of Brazil the same protection as President Trump, despite his expressions being of public interest. Ultimately, the exception gives social media platforms freedom to decide on a case-by-case basis.

Although the three companies have referred in one way or another to the defense of expressions of public interest on their platforms, none seems to have it entirely resolved. The use of the exceptions is also not included in the transparency reports that Facebook and YouTube regularly publish.

1. Facebook

"A handful of times a year, we leave up content that would otherwise violate our policies if the public interest value outweighs the risk of harm. Often, seeing speech from politicians is in the public interest, and in the same way that news outlets will report what a politician says, we think people should generally be able to see it for themselves on our platforms."[51] This is how Mark Zuckerberg explained the so-called newsworthiness exception.[52] This exception is not established in the community guidelines of Facebook but has been made known

H.R, case "Fontevecchia D'Amico vs. Argentina," judgment of November 29, 2011, § 61, retrieved from: https://corteidh.or.cr/docs/casos/articulos/seriec_238_esp.pdf, last access: March 5, 2021.

[51] Zuckerberg, Mark, Facebook, June 26, 2020, retrieved from: https://www.facebook.com/zuck/posts/10112048980882521, last access: March 5, 2021.

[52] Kaplan, Joel and Osofsky, Justin, "Input from Community and Partners on our Community Standards," Facebook, October 21, 2016, retrieved from: https://about.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards, last access: March 5, 2020, and Clegg, Nick, "Facebook, Elections and Political Speech," September 24, 2019, retrieved from: https://about.fb.com/news/2019/09/elections-and-political-speech, last access: March 5, 2021.

to the public through announcements on the company's blog or in public statements by its representatives.[53]

In the implementation of this exception, publications with news content that violate the Facebook rules, which are significant or important for the public interest may be kept up. In September 2019, Facebook reported that, as a general rule, posts by politicians are covered by this exception — which, however, does not apply to advertised content.

According to Facebook, to evaluate whether it is necessary to use the exception, the public interest value of the publication is weighed against the risk of harm, taking into account international human rights standards. If the damage is greater than the public's interest in knowing the content, Facebook would choose to remove it. According to the social media platform, when examining public interest, factors such as the circumstances of the country are taken into account, such as if an election is in progress or if the country is at war; the content of what was said, including whether it relates to governance or politics; and the political structure of the country, which means whether the country has a free press.

The limits of the exception have also been made public through statements by Facebook representatives. Nick Clegg, Facebook's vice president of global affairs and communications, said Facebook draws the line when it comes to posts that can lead to violence and harm in the real world. Mark Zuckerberg also asserted on his personal Facebook page that the exception does not apply to content that incites violence or seeks to suppress voters.[54] This time, Zuckerberg also reported that the company would begin to label the content it leaves online in the application of the exception and that it would allow people to share it in order to condemn it.[55]

However, the actual use arose some doubts. In September 2020, Facebook deleted for violation of its policy against incitement to violence a picture of a candidate (who was later elected) to the United States Congress for the state of Georgia where she posed with a rifle along with three photographs of Democratic politicians.[56] That month, Facebook deleted a post by a Louisiana congressman

[53] In a speech he gave at Georgetown University in 2019, Mark Zuckerberg stated that he did not believe that in a democracy it is okay for a private company to censor politicians or the news. Facebook, "Mark Zuckerberg Stands for Voice and Free Expression," October 17, 2019, retrieved from: https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression, last access: March 5, 2021.

[54] Zuckerberg, Facebook, *op. cit*.

[55] *Ibid*.

[56] Associated Press, "Georgia Candidate's Post Removed; Facebook Says It Violates Policy against Inciting Violence," Wabe,

promising the use of deadly force against protesters.[57] In contrast, the platform kept online a message from President Trump in which he argued, almost threateningly, that "when the looting starts, the shooting starts" also in the context of a public demonstration.[58]

COVID-19 has also shed light on inconsistencies in the use of policies: a video by Jair Bolsonaro in which he spoke in favor of the use of hydroxychloroquine in the treatment of the virus was removed from Facebook in March 2020,[59] but a Trump post with the same message was kept up.[60] However, another message in which Trump compares COVID-19 to the flu was removed from Facebook.[61]

2. Twitter

The public interest exception is more clearly established in the Twitter community rules.[62] According to these, a piece of content is of public interest "if it constitutes a direct contribution to the understanding or debate of a matter that concerns the whole public." Currently, the exceptions only apply to tweets from elected and government officials, and candidates for political office.[63] If a tweet covered by this exception is kept online, Twitter adds a warning or filter that provides context about the breach of the rules. This also makes the tweet not recommended by the Twitter algorithm and limits the ability of users to interact with what is posted.

September 4, 2020, retrieved from: https://www.wabe.org/georgia-candidates-post-removed-facebook-says-it-violates-policy-against-inciting-violence, last access: March 5, 2021.

[57] Associated Press, "Facebook Removes Congressman's Post over 'Incitement," ABC News, September 2, 2020, retrieved from: https://abcnews.go.com/Politics/wireStory/facebook-removes-congressmans-post-incitement-72778776, last access: March 5, 2021.

[58] "When the looting starts, the shooting starts." Trump, Donald, Facebook, May 28, 2020, retrieved from: https://www.facebook.com/DonaldTrump/posts/10164767134275725, last access: March 5, 2021.

[59] Constine, Josh, "Facebook Deletes Brazil President's Coronavirus Misinfo Post," Tech Crunch, March 30, 2020, retrieved from: https://techcrunch.com/2020/03/30/facebook-removes-bolsonaro-video, last access: March 5, 2021.

[60] Trump, Donald, Facebook, March 21, 2020, retrieved from: https://www.facebook.com/DonaldTrump/posts/10164254051615725, last access: March 5, 2020.

[61] Ingram, David, "Facebook Removes Trump Post that Compared Covid-19 to Flu," NBC News, October 6, 2020, retrieved from: https://www.nbcnews.com/tech/tech-news/facebook-removes-trump-post-compared-covid-19-flu-n1242277, last access: March 5, 2021.

[62] Twitter, "Acerca de las excepciones de interés público en Twitter" [About public-interest exceptions on Twitter], retrieved from: https://help.twitter.com/es/rules-and-policies/public-interest, last access: March 5, 2020. Two posts on the Twitter blog are also relevant: Twitter Safety, "Defining Public Interest on Twitter," June 27, 2019, retrieved from: https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html, last access: March 5, 2020, and Twitter Inc., "World Leaders on Twitter: Principles & Approach," October 15, 2019, retrieved from: https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html, last access: March 5, 2021.

[63] To apply the exception, the tweet has to be posted from a verified account that has at least 100,000 followers.

The platform weighs the public interest of the content against the possible risk and the severity of the damage. Unlike Facebook and YouTube, Twitter details in its rules the factors it considers to make its decisions:[64]

- There are violations where the exception is more likely to apply (for example, hate speech or harassment) and violations where the exception is less likely to apply and content is removed accordingly (such as in cases of terrorism, violence, or electoral integrity).
- Some criteria make the exception more likely to apply, such as when the tweet is directed at government officials or when it provides important context for ongoing geopolitical events. Furthermore, the exception is less likely to apply when, for example, the tweet includes a call to action.
- In no case are there exceptions made for multimedia content related to child sexual exploitation, non-consensual nudity, and violent sexual assault on victims.

In October 2019, Twitter addressed the use of the exception for world leaders, explaining that it is important to ensure people's right to know about their leaders and demand accountability.[65] Twitter explained that they are not entirely above its rules and that, in case of using the exception, a filter can be put in place. In November 2020, Jack Dorsey argued, in a hearing before the United States Congress, that Donald Trump would not have these protections when his term ends.[66] "If an account is suddenly not a world leader anymore, that particular policy goes away," said Dorsey. However, it is questionable to claim that world leader status is lost upon leaving the White House.

[64] Twitter publishes more detailed lists, here are some examples.

[65] Twitter Inc., "World Leaders on Twitter: Principles & Approach," *op. cit*.

[66] Hamilton, Isobel Asher, "Trump Will Lose his 'World Leader' Twitter Privileges on January 20, Jack Dorsey Confirms," Insider, November 18, 2020, retrieved from: https://www.businessinsider.com/donald-trump-twitter-account-lose-world-leader-protections-exemption-20-2020-11, last access: March 5, 2021.

Twitter applied the exception to this tweet by Donald Trump for violation of its rules on disinformation and COVID-19. Before someone can view the content, an interstitial warning appears.[67]

3. YouTube

The YouTube Community Guidelines do not properly establish a public interest exception. The only mention of such an exception is in a statement by Susan Wojcicki, YouTube's CEO, who argued that the platform could keep content posted by politicians that violates its rules: "When you have a public officer that is making information that is really important for their constituents to see, or for other global leaders to see, that is content that we would leave up because we think it's important for other people to see."[68]

After these words, a YouTube spokesperson stated that politicians are not treated differently by the platform, but are granted exceptions to certain kinds of speech with educational, documentary, scientific or artistic content — which YouTube groups under the acronym EDSA.[69] Among others, the EDSA principle appears in policies on harmful or dangerous content, violent or graphic content, incitement to hatred or violence, and, as mentioned before, in the policy on medical misinformation related to COVID-19. However, besides being an exception that allows infringing content to be kept online, the EDSA principle emphasizes the need to provide context to understand the intent of a video.[70]

[67] Twitter, October 11, 2020, retrieved from: https://twitter.com/realDonaldTrump/status/1315316071243476997, last access: March 5, 2020.

[68] Overly, Steven, "YouTube CEO: Politicians Can Break our Content Rules," Politico, September 25, 2019, retrieved from: https://www.politico.com/story/2019/09/25/youtube-ceo-politicians-break-content-rules-1510919, last access: March 5, 2021.

[69] *Ibid*.

[70] Google, "La importancia del contexto" [The importance of context], retrieved from: https://support.google.com/youtube/answer/6345162, last access: March 5, 2020.

V. New rules for advertising

All three platforms have taken various measures concerning paid advertising on their platforms. It is worth noting that, when it comes to advertised content, the reasoning is different: platforms generally do not allow any publication pending the identification of possible violations of their rules, as is the case with organic content. On the contrary, the contents to be advertised are previously presented by the contracting parties and approved by the platforms. That is why in these cases we do not speak of content removal but unauthorized content.

1. Unauthorized content: biosafety elements

All three platforms have taken various measures regarding advertised content that may affect the availability of biosafety items, even if the advertisements are not misinformative.[71] In February, before the declaration of the pandemic, Facebook banned ads that sought to create panic or denote urgency regarding supplies and products linked to COVID-19. In terms of fake content, the social media platform banned ads that guaranteed the cure or prevention of the virus. In March, Facebook banned sales of COVID-19 masks, hand sanitizers, disinfecting wipes, and test kits.[72] This latest ban has relaxed: since August 2020, Facebook allows the promotion of non-medical masks (subject to compliance with certain requirements),[73] hand sanitizer, and disinfecting wipes.[74]

Similarly, in April 2020 Twitter banned sensationalist or panic-inducing content and advertisements with inflated prices. Likewise, it prohibited the sale of masks and sanitizers and did not allow mention of vaccines, treatments, or test kits, except for information published by media outlets that the platform exempts under its policy of political advertisements.[75] Afterward, in August 2020, the ban was limited to medical masks.

---

[71] This document shows measures taken before November 30, 2020.

[72] Jin, "Keeping People Safe and Informed about the Coronavirus," *op. cit.*, "Banning Ads for Hand Sanitizer, Disinfecting Wipes and Covid-19 Testing Kits," updated on March 19, 2020, 2:18 PM.

[73] Leathern, Rob, "Allowing the Promotion of Non-Medical Masks on Facebook," Facebook, June 10, 2020, retrieved from: https://www.facebook.com/business/news/allowing-the-promotion-of-non-medical-masks-on-facebook, last access: March 5, 2021.

[74] Jin, "Keeping People Safe and Informed about the Coronavirus," *op. cit.*, updated on August 19, 2020, 10:05 AM.

[75] Twitter, "Contenido de carácter político" [Political Content Policy], retrieved from: https://business.twitter.com/es/help/ads-policies/ads-content-policies/political-content.html, last access: March 5, 2021.

Finally, like Facebook, Google restricts the sales of face masks required by health workers, to avoid shortages of supplies, and allows only the sale of exclusively cloth face masks.[76] The company also reported that it is taking steps to prevent artificial price surges.

	Mentions[77]	Sales	Other
FB		Medical masks	Panic/urgency
		Hand sanitizer	Cure/prevention of the virus
		Disinfectant wipes	
		Test kits	
TW	Vaccines	Face masks	Panic/ sensationalist
	Treatments	Hand sanitizer	Artificial prices
	Test kits		
YT		Medical masks	Artificial prices

2. Unauthorized content: terms related to COVID-19

In addition to the restrictions related to biosafety elements, the platforms took other measures regarding advertised content. Both Twitter and Google started the pandemic with a blanket ban on any advertising that used words related to COVID-19. This being the case, no publications were allowed on matters of public interest or that could be useful to users — such as donation channels or health insurance. According to Google, this was decided by applying its inappropriate content policy, which was in place before the pandemic.[78] Under this policy, advertisements that potentially profit from a "sensitive event" such as a natural disaster or conflict are not allowed. With the pandemic, Google updated the policy to include "public health emergencies" as a type of sensitive event.[79]

[76] Google, "Actualizaciones de la política de Google Ads sobre la enfermedad del coronavirus (covid-19)" [Coronavirus disease (COVID-19) Google Ads policy updates], retrieved from: https://support.google.com/google-ads/answer/9811449?hl=es-419, last access: March 5, 2021.

[77] November 30, 2020 update.

[78] Pichai, Sundar, "Coronavirus: cómo estamos ayudando" [Coronavirus disease: How are we helping], Google, March 6, 2020, retrieved from: https://blog.google/inside-google/company-announcements/coronavirus-covid19-como-estamos-ayudando, last access: March 5, 2021, and Google, "Contenido inapropiado" [Inappropriate content], retrieved from: https://support.google.com/adspolicy/answer/6015406?hl=es-419, last access: March 5, 2020.

[79] Google, "Actualizaciones de la política de Google Ads sobre la enfermedad del coronavirus (covid-19)" [Coronavirus disease (COVID-19) Google Ads policy updates], *op. cit*.

Google's decision was strongly questioned by members of the United States Democratic Party.[80] Since the ban had exceptions for government bodies, the Donald Trump administration was allowed to post announcements concerning its response to the pandemic, while Democrats did not have the opportunity to present critical publicity of the government's actions towards the crisis.

In April, these companies introduced some changes.[81] In addition to banning the sales of biosafety items, Twitter banned tasteless advertising references to COVID-19.[82] Google, for its part, reported that it would allow advertisements from "health care providers, governmental, non-governmental and intergovernmental organizations, advertisers who publish verified electoral messages, and privately managed accounts with a history of compliance with policies that want to share relevant information with the public."[83]

In December 2020, anticipating the start of vaccination against COVID-19 worldwide, Facebook announced that it would ban advertised content where vaccination kits were sold or accelerated access to the vaccine was promoted. Facebook also bans ads that claim the vaccine is a cure for the virus.[84]

3. Elimination of the category "pseudoscience"

In April 2020, the media outlet The Markup showed how it was possible to hire advertising directed at people who, according to Facebook, were interested in "pseudoscience."[85] This category encompasses more than 78 million people. After the publication of The Markup, Facebook quietly eliminated this category as

[80] Birnbaum, Emily, "Democrats Say Google's Covid-19 Ad Ban Is a Gift to Donald Trump," Protocol, March 31, 2020, retrieved from: https://www.protocol.com/google-coronavirus-ad-ban-democrats, last access: March 5, 2021.

[81] Fischer, Sara, "Twitter Lifts Coronavirus Ad Ban," Axios, April 3, 2020, retrieved from: https://www.axios.com/twitter-coronavirus-ad-ban-eb9c8946-d90b-4bb2-be29-9fba531e44d8.html, last access: March 5, 2020, and Birnbaum, Emily, "Google Revises Covid-19 Ad Ban after Backlash", Protocol, April 2, 2020, retrieved from: https://www.protocol.com/google-coronavirus-ad-ban-reverse, last access: March 5, 2020.

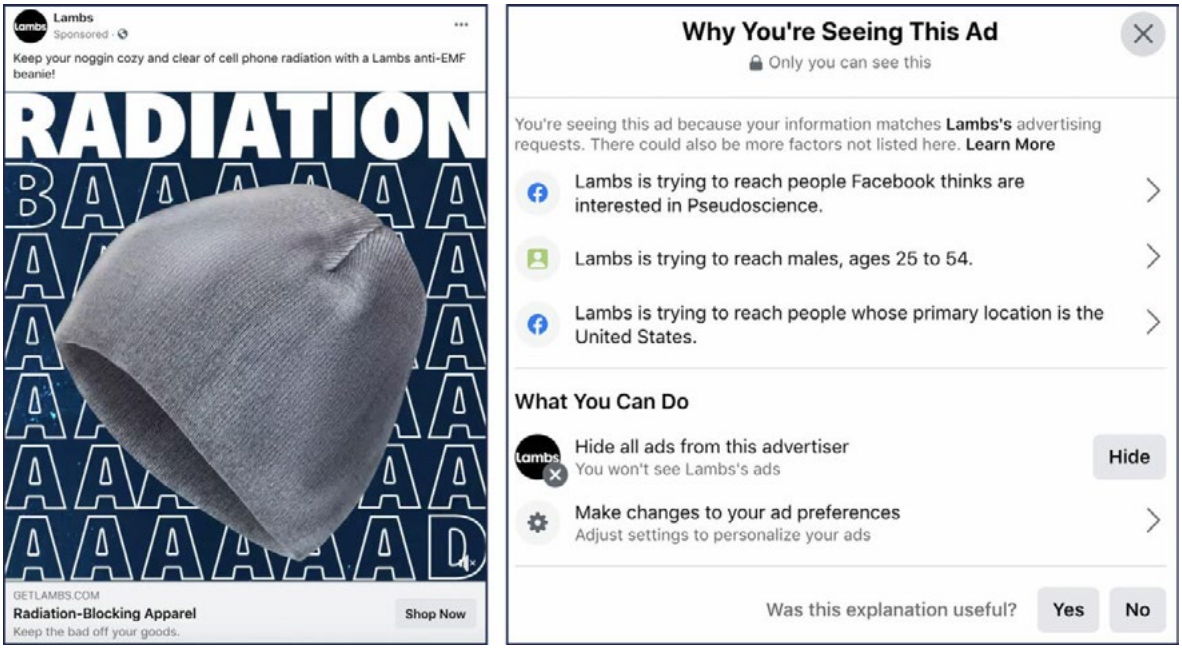
[82] Ad Exchanger, "Google, Twitter Lift Coronavirus Ad Ban; Facebook Offers Grants to SMBs," April 6, 2020, retrieved from: https://www.adexchanger.com/ad-exchange-news/monday-06042020, last access: March 5, 2021.

[83] Google, "Actualizaciones de la política de Google Ads sobre la enfermedad del coronavirus (covid-19)" [Coronavirus disease (COVID-19) Google Ads policy updates], *op. cit*.

[84] Jin, "Keeping People Safe and Informed about the Coronavirus," *op. cit*.

[85] Sankin, Aaron, "Want to Find a Misinformed Public? Facebook's already Done It," The Markup, April 23, 2020, retrieved from: https://themarkup.org/coronavirus/2020/04/23/want-to-find-a-misinformed-public-facebooks-already-done-it, last access: March 5, 2021.

an option for the purchase of targeted ads, which could ultimately impact publicity around COVID-19.[86]



The Markup found an ad for a hat that claims to protect people's heads from cellular radiation. Concerns about electromagnetic radiation emitted by the 5G mobile phone infrastructure are part of the conspiracy theories related to the Coronavirus disease. By clicking "Why are you seeing this ad?" Facebook showed that the advertising company wanted to reach people who may be interested in pseudoscience.3

4. Demonetization

Google also took a step that affected another group of players in its ad ecosystem: the creators of the YouTube partner program. As mentioned above, YouTube gives a portion of the advertising revenue to the accounts that are part of this program.[88] However, at the beginning of the pandemic, the company announced that it would not show advertising (what is known as "demonetizing") in videos focused on COVID-19.[89] This was also done by extending its policy

[86] Culliford, Elizabeth, "Facebook Gets Rid of 'Pseudoscience' Ad-targeting Category," Reuters, April 23, 2020, retrieved from: https://uk.reuters.com/article/health-coronavirus-facebook-ads/facebook-gets-rid-of-pseudoscience-ad-targeting-category-idUKL2N2CB1D6?feedType=RSS&feedName=technology-media-telco-SP, last access: March 5, 2021.

[87] Sankin, "Want to Find a Misinformed Public? Facebook's already Done It," op. cit.

[88] At least 1,000 subscribers and 4,000 hours of playback are required to join. Google, "Descripción general y elegibilidad del programa de socios de YouTube" [YouTube Partner Program Overview and Eligibility], *op. cit*.

[89] Barca, Kamila, "Youtube ha desmonetizado los videos de influencers que hablan sobre el coronavirus" [YouTube has demonetized the videos of influencers who talk about the Coronavirus disease], Business Insider, February 18, 2020, retrieved from:

related to "sensitive events." This measure was also relaxed. In March, YouTube announced on its blog that it would allow ads for videos discussing COVID-19 on a limited number of channels.[90]

VI. Conclusions

It will be a long time before we understand the impact that the Coronavirus pandemic has had on humanity and, in particular, the effects of this global event on the exercise of freedom of expression on the Internet. However, a consequence that we can predict in this sector is the dramatic change in the moderation of content on social media.

It would be wrong to assert, in any case, that this new scenario occurred solely on account of COVID-19. The platforms had been facing political, regulatory, and social pressure of all kinds to intervene more decisively in the problematic content of users. For example, 2020 was also the year of the presidential election in the United States: the call to action for social media was to avoid repeating the situation of the 2016 elections.

This pandemic emerged in a globalized society in a complex digital environment, where interconnection offers both solutions and problems: false rumors and miracle cures circulated through the same networks where the first alerts and public health measures began to spread. In this context, services such as Facebook, YouTube, or Twitter — which were already facing several critics regarding their responsibility in shaping the public debate — focused on saving their own skin while preserving their business model. The platforms' main objective was to avoid at all costs becoming a vector of misinformation and mistrust as contagious as the Coronavirus disease.

In the following months, there will be a debate on the solutions that these platforms opted for. For now, bearing in mind the exceptionality of the problem, without erasing the previous work, we offer some reflections and conclusions:

- **The pandemic landed on social media amid many questions about content moderation.** Although this exacerbated the problem, there was already a

https://www.businessinsider.es/youtube-ha-desmonetizado-todos-videos-coronavirus-583723, last access: March 5, 2021.

[90] Wojcicki, Susan, "Coronavirus: An Update on Creator Support and Resources," Official YouTube Blog, March 11, 2020, retrieved from: https://blog.youtube/inside-youtube/coronavirus-update-on-creator-support, last access: March 5, 2021.

general state of dispersion, ambiguity, and inconsistency about the community regulations to which users are subject, where they are and how they are implemented. Many rules were announced in blog posts or the news without ever being formally incorporated.

- **What is the nature of the Community rules?** The objective of this work was not to make a hermeneutical analysis of these texts. However, there is a fundamental question about the nature of these rules. In some cases, the platforms seem to postulate the rules in an exhaustive way and in others as the enunciation or formulation of a principle subject to interpretation and development. In one way or another, platforms have the first and last word.
- **The perfect storm of COVID-19 and the US presidential campaign.** The main goal of social media facing the 2020 elections in the United States was to avoid the use of the platforms that occurred in 2016: access to databases, segmentation of advertising, and joint operations enabled a manipulated, highly polarized, and, in essence, uninformed debate. With the arrival of the Coronavirus disease, the health and electoral emergencies concurred, which highlighted the limitations that these companies face in controlling political actors who capitalize on disinformation strategies — starting with Donald Trump.
- **The difficulty in appointing reliable or official authorities.** The pandemic also incited a crisis of confidence. With the best intentions, the platforms tried to give more prominence to authoritative public health sources. However, the process used to show confidence was not transparent, nor was there a way to resolve the obvious contradictions between the different authorities in a country — as was the case in Brazil.
- **Ambiguity regarding public interest.** The difficulty of establishing clear criteria to define the public interest exception was more evident at this time. During this crisis, many political leaders spread problematic messages — such as miracle cures — that, however, had a connotation of public interest. In the end, what seemed to prevail in the application of the exception was avoiding a public image crisis for the platforms.
- **Moderation decisions during the political debate.** The way in which the public interest exception was implemented regarding President Donald Trump highlights the political moment in which that decision had to be made and — ultimately, the vision of the platforms — which translated into different ways of moderating content. Twitter, which began to set limits to Trump,

was the first to label his tweets and, although it applied the public interest exception, with the measure it substantially reduced the possibility of interacting with the offending content. Facebook, for its part, was more reluctant to mess with the president, but by using the exception and labeling content it still allows interactions.

- **Between form and substance.** Each platform has emphasized differently how it presents community standards in this emergency. While Facebook offers more details in substance, Twitter emphasizes processes. Among the three, YouTube usually finds a balance between both points.
- **More clarity in the rules for advertisers.** It seems that the platforms have been more concerned with clearly delineating the rules that advertisers must follow during the pandemic than with the community rules for users.

The decisive and imperfect intervention of the platforms during the pandemic gives way to a new instance in the discussion about content moderation. It is possible that as normality returns, some rules will be reversed or nuanced — as, indeed, it has already happened — but it is hard to think that these intermediaries can return to the previous vision (more focused, for example, on inauthentic activity than in content) and to the position of not wanting to be arbiters of the truth.

2020 for social media also coincided with the entry into operation of Facebook's oversight board and with several civil society initiatives that try to influence the debate on the liability of intermediaries in shaping the digital public debate: among others, the boycott against advertisers "Stop Hate for Profit";[91] the "Change the Terms" campaign to reduce online hate;[92] updating the Santa Clara Principles[93] on transparency and accountability in content moderation; and, in Latin America, the project led by Observacom on standards for the regulation of platforms.[94]

These initiatives have wanted to focus the conversation on key questions about

[91] Stop Hate for Profit, retrieved from: https://www.stophateforprofit.org, last access: March 5, 2021.

[92] Change the Terms, retrieved from: https://www.changetheterms.org, last access: March 5, 2020.

[93] Santa Clara Principles, retrieved from: https://santaclaraprinciples.org, last access: March 5, 2021.

[94] Libertad de Expresión y Plataformas de Internet, "Estándares para una regulación democrática de las grandes plataformas que garantice la libertad de expresión en línea y una Internet libre y abierta" [Standards for a democratic regulation of large platforms that guarantees freedom of expression online and a free and open Internet], July, 2020, retrieved from: https://www.observacom.org/wp-content/uploads/2020/09/Estandares-para-una-regulacion-democratica-de-las-grandes-plataformas.pdf, last access: March 5, 2021.

community guidelines: what standards should guide them? How to make them compatible with human rights? How to achieve accountability and some form of "due process"? How to guarantee the validity of freedom of expression? While these are fundamental issues on the human rights agenda in the digital age, solving these questions would not fully solve the practical dilemma of content moderation — issues such as scale, consistency, or timing — or economic, social, and political problems that exist in social media and that the pandemic exacerbated. Perhaps the exceptionality of the year 2020 applies also to this question: it is time to reinvent this conversation.

Annexes

The following tables summarize the measures taken by Facebook, Twitter, and YouTube between February and November 2020, in accordance with what was explained in previous sections. Content is divided into organic content and advertised content. Additionally, for each action, it is specified if, as far as it was possible to verify, the measure was taken when applying a previously established community regulation, if it was communicated through the company's blog or if the change was incorporated directly into the text of the community guidelines.

In general terms, the measures taken consist of:

- Identifying content marked as false by fact-checkers or health authorities, to eliminate it, label it or reduce its circulation.
- Sending notifications to people who plan to share or who have been in contact with content marked as false.
- Restricting paid advertisements and organic publications pretending to sell certain biosafety elements (such as masks or hand sanitizer), seeking to create panic, or which claim to cure or prevent the virus.

Facebook				
Type of content	Measure	**Application of a standard already in force?**	Measure announced on the blog?	**Change incorporated as a new community standard?**
Organic	Eliminating content contrary to health authorities	Debatable	Yes	No
	Identification of duplicates	Unclear	Yes	No
	Prohibiting organic sales of certain biosafety items	No	Yes	No
	Labels and restricting dissemination of content disputed by fact-checkers	Yes	Yes	N/A
	Notifications of exposure to disinformation content	No	Yes	No
	[Instagram] Blocking or restricting the use of hashtags	No	Yes	No
	[Instagram] Removal of recommendations related to COVID-19	No	Yes	No
	[Instagram] Disabling the option to search for augmented reality effects related to COVID-19	No	Yes	No
Advertising	Restricting the sale of certain biosafety items and advertisements that seek to create panic or that guarantee the cure or prevention of the virus	No	Yes	Yes
	Eliminating the category “pseudoscience”	N/A	No	No
	Restricting the sale of vaccine kits or the promotion of accelerated access to them	No	Yes	Yes

Twitter				
Type of content	**Measure**	**Application of a standard already in force?**	**Measure announced on the blog?**	**Change incorporated as a new community standard?**
Organic	Eliminating tweets contrary to health authorities	No	Yes	No
	Eliminating misleading information with a propensity for severe harm	No	Yes	No
	Filters for misleading information with moderate propensity to harm or controversial claims with severe propensity to harm	No	Yes	No
	Labels can be used in cases of controversial claims with moderate propensity to harm and in some cases of unverified claims with severe propensity to harm	No	Yes	No
Advertising	Prohibiting organic sales of certain biosafety items	No	Yes	Yes
	General ban on words related to COVID-19	No	Yes	Yes

YouTube				
Type of content	**Measure**	**Application of a standard already in force?**	**Measure announced on the blog?**	**Change incorporated as a new community standard?**
Organic	Prohibiting recommendations for dangerous substances or harmful treatments	Yes	N/A	N/A
	Prohibiting content about COVID-19 that implies a serious risk of flagrant harm	No	Yes	Yes
	Prohibiting content contrary to health authorities	No	Yes	Yes
Advertising	Prohibiting organic sales of biosafety items	No	Yes	Yes
	General ban on ads with words related to COVID-19	Yes, with clarifications	Yes	N/A
	Demonetization of content focused on COVID-19	No	No	No