CELE

# Fake news on the Internet: actions and reactions of three platforms

*February 2021*

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

UP
Universidad
de Palermo

# Fake news on the Internet: actions and reactions of three platforms[1*]

In the past four years, online misinformation has become a seemingly novel threat to Western democracies. This threat surfaces in a time of anxiety about the future: democratic systems appear to be going through a crisis as a result of the consolidation of political options that, even if they do not openly reject democratic procedures and formalities, erode them from the inside. It is in this context that the phenomenon of disinformation must be addressed.

In general, anxiety about democracy provokes a reaction. In the case of disinformation, the reaction came from the states, researchers, civil society, and the large Internet platforms where misinformation seems to spread. This paper explores the actions of the latter in the face of the claims of the former and frames the question during a key moment about the future of the Internet.

The first section presents a general conceptualization of the phenomenon of disinformation on the Internet and how it re-surfaced on the public agenda. Likewise, although the dissemination of false information to deceive the public has always been part of the playbook of different state and parastatal actors, the efforts to influence the public debate in the United States during the 2016 electoral campaign introduced a period in which this old practice acquired new characteristics, which were replicated and mutated in subsequent electoral processes such as those of the United Kingdom (2016), Colombia (2016), Mexico (2018) and Brazil (2018). This time-based aspect of the phenomenon is relevant for its definition since it links the reactions of the large Internet platforms with a specific crisis, which is also fueled by the widespread pessimistic turn regarding the Internet's ability to have a positive impact in the future of democracy.

The second section presents a general outline of the main actions taken by platforms in the matter of disinformation, from late 2016 to late 2020. The objective of this temporal approximation is twofold: to identify public actions, especially in the context of Latin America, and verify their degree of implementation. The slow increase of announced changes and responses to the phenomenon suggests

[1] This document was written by Ramiro Álvarez Ugarte and Agustina Del Campo.

an industry under pressure, trying to find satisfactory answers to a growing demand for the moderation of problematic content.

The third section discusses the findings and postulates the existence of an "aporia" at the center of the problem. The pessimistic turn regarding the Internet and the process of concentration of traffic on large platforms has raised new questions about the current legal solution concerning the flow of information online. The platforms' actions suggest that they are moving towards a public forum model, where the principles and criteria they use to define which speech is acceptable or not is entirely under their control. This model calls into question the principle of non-liability of intermediaries for third-party content, which until now was the predominant legal solution.

## A. Disinformation as a current problem of democracy

One possible definition of disinformation is "the massive dissemination of false information (a) with the intention of misleading the public and (b) knowing that it is false."[2] Using this definition as a starting point allows us to see that we are not facing a particularly new phenomenon: various actors have tried to manipulate the population through false information throughout history.[3] Perhaps for this reason, to analyze the phenomenon it seems more relevant to limit it temporarily than conceptually.

Indeed, *disinformation* emerges as a recent phenomenon in the framework of electoral processes characterized by an impoverished public debate in part thanks to the massive dissemination of false information, especially after the 2016 United States electoral campaign that led Donald J. Trump to the presiden-

[2] IACHR (*Guía para garantizar la libertad de expresión frente a la desinformación deliberada en contextos electorales* [Guide to guarantee freedom of expression regarding deliberate misinformation in electoral contexts.]) Office of the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights, Washington D.C., October 17, 2019.), 3. See also C. Botero Ignacio Álvarez, Eduardo Bertoni, Catalina Botero, Edison Lanza (eds.) («*La Regulación Estatal De Las Llamadas "noticias Falsas" Desde La Perspectiva Del Derecho A La Libertad De Expresión*» [State Regulation Of The So-called" Fake News From The Perspective Of The Right To Freedom Of Expression], in *Libertad de expresión: A 30 años de la Opinión Consultiva sobre la colegiación obligatoria de periodistas*[Freedom of expression: 30 years after the Advisory Opinion on the compulsory membership in an association prescribed by law for the practice of journalism], 1st, Inter-American Commission on Human Rights, Washington D.C., 2017, (OAS. Official documents; OEA / Ser.D / XV.18).), 69; M. Verstraete; OF Bambauer; JR Bambauer ("Identifying and Countering Fake News." Social Science Research Network, Rochester, NY. ID 3007971. August 1, 2017.)

[3] Several studies have made this point: see L. Bounegru; J. Gray; T. Venturini; M. Mauri («A Field Guide To"Fake News" And Other Information Disorders». Public Data Lab & First Draft, Amsterdam. 2017.), 6; R. Darnton ("The True History Of Fake News," The New York Review of Books, 2/13/2017, retrieved from https://www.nybooks.com/daily/2017/02/13/the-true-history-of-fake-news/ Last access: 1/March/2019.)

cy.[4] Since then, in various elections, disinformation appeared as a more or less serious problem: the referendums for the permanence of the United Kingdom in the European Union and peace in Colombia in 2016, and the 2018 presidential elections of Brazil and Mexico are good examples of this.

Various observers conclude that disinformation is a serious threat to the future of democratic systems since electoral processes are the main mechanism through which governments acquire legitimacy. This perception, shared by some states, caused various kinds of reactions — regulatory, awareness-raising, or self-regulation — that increased pressure on large Internet platforms, perceived as "facilitators" of the phenomenon.[5] State pressure, together with the damaged reputation produced by questionable business practices in the handling of users' personal data, is the main cause of the actions announced or adopted by the platforms in recent years, which this report documents. But understanding these reactions requires awareness of the historical nature of the phenomenon. *Disinformation* happens in two distinct but related crises: the crisis of democracy; and the crisis of the optimistic Internet model that justified — until now — the main regulatory approaches to how information circulates on this network.

## *1. Internet and the age of technological optimism*

As a decentralized network, the Internet was presented from the beginning as the ideal space for the free circulation of information and ideas of all kinds. [6] The creation of the network, and — particularly — its popularization from the second half of the 1990s, radically changed the way information circulates: from highly centralized models controlled by powerful actors to a decentralized model where, with very low investment in technology, each one of the users could

[4] Cf. C. Cortés; L. Isaza, «*Noticias falsas en Internet: La estrategia para combatir la desinformación*» [Fake news on the Internet: The strategy to combat disinformation]. Centro de Estudios para la Libertad de Expresión, Buenos Aires. December 2017. Page 2 (referencing the infamous «Pizzagate»).

[5] One of these reactions is the German law of June 30, 2017, known as Netzwerk Durchsetzungsgesetz or NetzDG, which imposes various obligations on platforms concerning numerous instances of illegal content. The law was criticized by the OSCE Special Rapporteur on Freedom of the Media and by Human Rights Watch. Cf. OSCE (*Osce Representative On Freedom Of The Media Warns Germany Social Networks Law Could Have Disproportionate Effect*, OSCE, 10/04/2017, retrieved from https://www.osce.org/fom/347651 Last access: 4/March/2019.);H. R. W. |. 350. F. Avenue; 34th. F. |. N. York; N. 10118. U. |. t 1.212.290.4700 (*Germany: Flawed Social Media Law*, Human Rights Watch, 02/14/2018, retrieved from https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law Last access: January/ 16/2019.). France enacted a similar rule in November 2018, specifically aimed at combating "information manipulation."

[6] On this description, see eg Y. Benkler, The Wealth Of Networks: How Social Production Transforms Markets And Freedom, Yale University Press, New Haven [Conn.], 2006 (which reflects this "optimistic" view on the network).

produce their own content and disseminate it under similar conditions of large producers of traditional content.[7]

This initial promise of decentralization generated some optimism regarding the Internet and its influence on the future of democracy because some basic characteristics of the network design tend to favor the flow of information over its control.[8] This design strengthens the resilience of the network and does not depend on nodal points of control to function.[9]

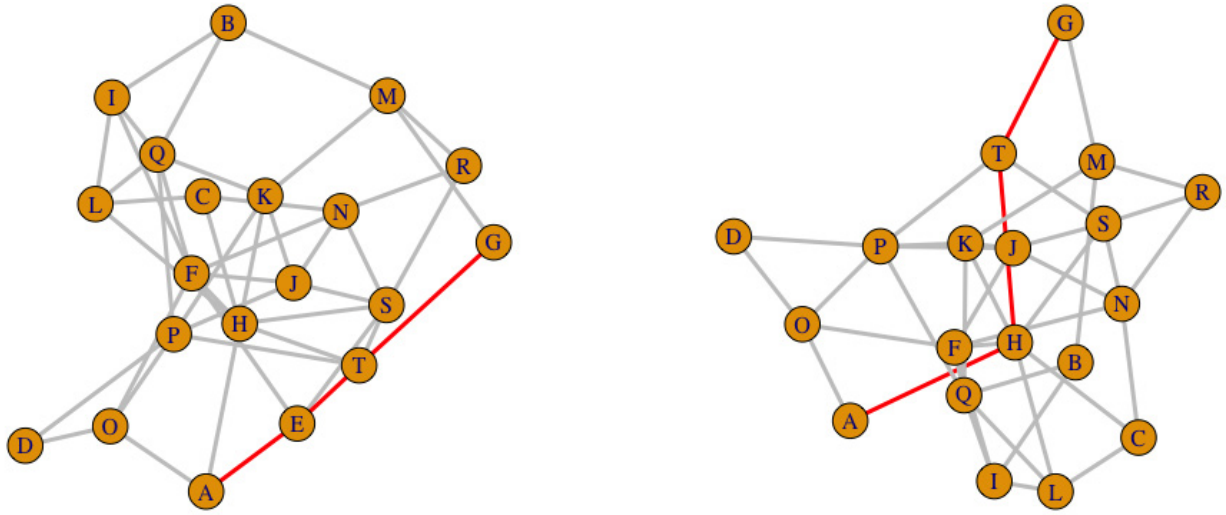


**Figure 1**

(a) Decentralized random network of 20 nodes

(b) Decentralized random network of 19 nodes (-E)

[7] Cf. Y. Benkler (*Ibid.*), Chap. 5 (where to put the famous home film production The Jedi Saga, for example). Much of the pioneering literature on law and the Internet pointed to this potentially beneficial dimension of the Internet. See also e.g., L. Lessig (*Code*, Version 2.0, Basic Books, New York, 2006.); E. Moglen ("The Invisible Barbecue," Columbia Law Review, vol. 97, 1997.) ("Any individual can reach through the use of network media such as the World Wide Web an audience far larger than the owner of several small television stations, and at no cost. The form of communication achieved is no doubt different, and it is obvious that the individual UseNet news post has not achieved parity with a television advertisement for an athletic shoe. But the metaphor of the broadcaster and the consumer conditions is to accept the maximum possible inequality as a natural necessity: a few talk and the rest merely listen. Spectrum need not be allocated around this conception; the historical justifications for doing so are among the forms of reasoning vitiated by the rise of complementary carriage technologies and the resulting death of 'scarcity rationales'")

[8] Cf. L. Lessig (*Code*, cit.), ix-x; an optimism that — at times — led to what Eugene Morozov rightly called cyber-utopianism. Cf. E. Morozov (*The Net Delusion: The Dark Side Of Internet Freedom*, 1st ed, Public Affairs, New York, 2011.). *See also* J. Zittrain (The Future Of The Internet And How To Stop It, Yale University Press, New Haven [Conn.], 2008.), 30-35 (describing the design principles of the original Internet)

[9] The end-to-end (e2e) principle on which the first version of the Internet was built delegates most of the work to the ends of the network: the network must be "as simple as possible, and its intelligence must be on the extremes or endpoints of the network." This principle explains why when speaking of the Internet we speak of a *decentralized* network - the network could have been designed differently, e.g. by establishing nodal points to which each user must connect for access. But it was not like that."

Figure 1 represents a decentralized network, in which nodes are connected to each other in various ways. For efficiency, a network of this nature seeks to connect the points farthest from each other through the shortest possible steps. In Figure 1.a the link between node A and node G is highlighted: the shortest possible path takes three steps (A → E → T → G). If we remove the first node of that section ("E", in Figure 1), we see that the node quickly finds another equally efficient path (A → H → T → G). In a less decentralized network, in which some nodes play the role of brokers (such as, for example, node "K" in Figure 2), the connection of nodes D → G becomes impossible.

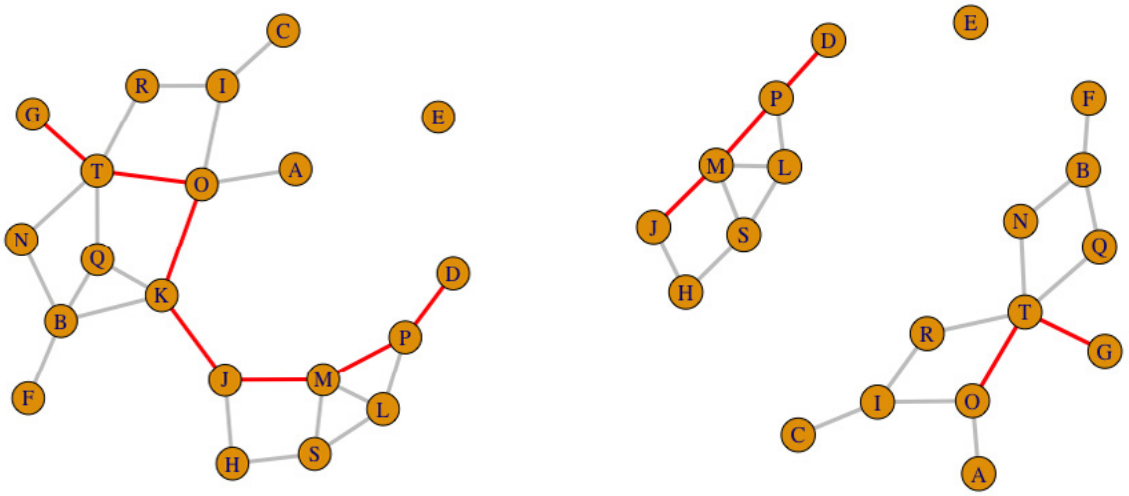


Figure 2

(a) (b)

When the Internet became more popular, some of the assumptions of the original model were questioned. Thus, for example, towards the end of the 1990s, the security risks posed by a decentralized network became apparent.[10] However, the optimistic outlook persisted and was evident in the confidence that the Internet could be a democratizing tool in closed societies since its decentralized nature would allow it to corrode the structures that permit totalitarian governments to control their citizens.[11] In this context, legal protections soon emerged for intermediary actors that facilitated access to content produced by third parties. Section 230 of the Communications Decency Act of 1996 established, for example,

[10] On this point, see especially Ibid., Chap. 3.

[11] E. Morozov, The net delusion, cit., *passim*.

the principle of non-liability of intermediaries to safeguard the original architectural design of the network. Who would risk fulfilling this role of intermediation if they could be held responsible for the information whose circulation it facilitates? This model was later adopted in other countries, through legislation, jurisprudence, or soft law.[12]

## 2. The age of pessimism

The optimistic views on the Internet had detractors from the start. Many authors saw how certain trends in the development of the network could lead to an Internet that is more closed, regulated, and subject to the control of state and private actors. Thus, Lawrence Lessig pointed out in 1999 how the pressures of commerce and the states would tend to regulate the Internet more rigorously, both through legislation and in requirements in the code used to design the network's architecture.[13] Similarly, Jonathan Zittrain predicted how security concerns could lead to the return of technology models that the Internet seemed to have left behind (manufacturer-controlled "home appliances" that do only what manufacturers want, e.g., cell phones).[14] A modicum of caution was emerging: the Internet was not necessarily beneficial for democracy. Eugeny Morozov was one of the first to point out how authoritarian governments were learning to use the Internet to increase their control over the population.[15] These warnings were the sign of the start of a "pessimistic turn," which was undoubtedly fueled by the mutation of the original architecture of the network.

During the expansion of the Internet, various intermediary actors in the circulation of information emerged and became more and more powerful. Thus, Google (1998), for example, achieved a dominant position in the search engine market and became a key player in accessing previously unknown information. YouTube (2005) achieved something similar in the online video market, due to increased connection speeds. Facebook (2004) and Twitter (2006) made it possible to connect people by replicating on the Internet physical social networks and creating new ones. These companies began to concentrate a large part of the traffic and thus a process

[12] *See,* e.g., the Manila Principles, in https://www.manilaprinciples.org/es. For an opposite point of view, see J. Kosseff, The Twenty-six Words That Created The Internet, Cornell University Press, Ithaca [New York], 2019, Chap. 7 (where the author points out that section 230 is an example of the "exceptionalism" of the United States, and argues that other countries have adopted more moderate versions of that provision.

[13] L. Lessig, *Code*, cit.

[14] J. Zittrain, The Future Of The Internet And How To Stop It, cit., Chap. 1.

[15] E. Morozov, The net delusion, cit.

of re-centralization of the network took place.[16] These actors expanded, acquired new services — in many cases, direct competitors — and consolidated their dominant positions thanks to the network's effects that generate incentives to participate in these centralized channels (and increasing costs of "disconnection").[17]

Within the framework of this re-centralization, these intermediary companies developed terms of services and community guidelines that operate as *de facto* regulations of the information that is considered acceptable in each of these platforms. Furthermore, they more decisively embraced their role of moderators when they introduced timelines based on algorithms that — supposedly — seek to bring users more "relevant" content, determined by consumption patterns that the platforms record and exploit.[18]

These behaviors underlined the essential role these companies play in the events and led many legislators around the world to wonder why representative bodies of democracy cannot intervene in the definition of these criteria.[19] This power, which was wielded erratically but steadily by intermediaries, over time, was

---

[16] See S. Hubbard («Fake News Is A Real Antitrust Problem», 12/2017, retrieved from https://www.competitionpolicyinternational.com/fake-news-is-a-real-antitrust-problem/ Last access: January/ 16/19), 2 ("Two corporations have an outsized control on the flow of information worldwide. Google accounts for roughly 80 percent of global Internet searches, and its search market share exceeds 90 percent in most European countries. Facebook dwarfs all other social networks, with two billion active monthly users"). See also Observacom («Ott Regulation. Key Points For The Democratic Regulation Of "over-the-top" Services So As To Ensure A Free And Open Internet And The Full Exercise Of Digital Rights And Freedom Of Expression». Observacom. September 2017.), 3:

"In a scenario centralized by the traditional media, it was clear that the market on its own did not guarantee the fundamental diversity, pluralism, and freedom of expression needed by democracy. With the emergence of the Internet, it seemed that part of the rationality that gave meaning and foundation to democratic regulation might have been lost. Some important players in the digital ecosystem claim that regulation of the Internet is not only dangerous but should not exist, as it is no longer necessary or possible. However, after the initial phase of more decentralized and open network operation, new bottlenecks formed and the Internet embarked on a growing centralization among just a few actors of the digital ecosystem that has affected its potential to serve all of humanity: this was underlined by the creator of the World Wide Web, Tim Berners Lee. The trend towards concentration and threats to freedom of expression on the Internet show that diversity and pluralism—and even the notion of an open and free Internet—need regulatory guarantees so that they can be maintained as values and paradigms of modern digital communications."

[17] On this point, *see* K. S. Baran; K. J. Fietkiewicz; W. G. Stock («Monopolies On Social Network Services (sns) Markets And Competition Law», ISI, 2015.) (Describing the network effects that lead to monopolies or quasi-monopolies in social media markets). See also E. Morozov (*The net delusion*, cit.), 217 ("While activists can minimize their exposure to intermediaries by setting up their own independent sites, chances are that their efforts may not receive the kind of global attention that they might on Facebook or YouTube").

[18] This model is different from the one that was based, purely and exclusively, on user choices: from the "portals" that manually curated information to the "RSS feeds" model that brought information easily accessible but based on the "express" subscription of users. On this change, see M.A. DeVito; D. Gergle; J. Birnholtz ("Algorithms Ruin Everything": #riptwitter, Folk Theories, And Resistance To Algorithmic Change In Social Media," CHI '17: CHI Conference on Human Factors in Computing Systems, Denver Colorado USA, May 2, 2017, retrieved from https://dl.acm.org/doi/10.1145/3025453.3025659)

[19] On this point, the trends are of course contradictory: while some want to toughen the platforms' criteria to deal with various problems (disinformation, defamation, copyright violations, hate speech, etc.) others want to protect the users of decisions that are arbitrary or based on obscure criteria.

accompanied by growing demand from various actors for greater liability and accountability.[20] And this, in turn, has pushed platforms to do much more than simply "facilitate" the flow of information: they have become "curators" of that content, both for their own reasons (such as developing content recommendation algorithms) and external pressure.[21]

Finally, this concentration process was facilitated by the expansion of a business model that provides these services for "free" under a traditional advertising system that makes use of the analysis and personal data of citizens to deliver highly personalized messages.[22] This model has been the subject of criticisms in recent years, partly driven by some scandals that exposed to what extent privacy is compromised in this model, including the Edward Snowden revelations in 2013 or the Cambridge Analytica scandal and the use of personal data obtained from Facebook for political purposes in 2018.[23] This advertising model is not necessarily in crisis, but criticism in recent years has contributed to the "pessimistic" turn.

## *3. Disinformation as a problem*

Since the 2016 US presidential election, the main intermediary platforms appear to be on the defensive. On a regular and somewhat chaotic basis, they have since announced and implemented changes to their platforms aimed at combating the phenomenon of disinformation. They have also been summoned to testify at informational hearings before legislative bodies around the world. Furthermore, they

[20] On this issue, the literature on accountability and tech companies seems to have increased exponentially in recent years. *See* S. Pearson; N. Wainwright («An Interdisciplinary Approach To Accountability For Future Internet Service Provision», International Journal of Trust Management in Computing and Communications, vol. 1, 1, 2013, retrieved from https://www.inderscienceonline.com/doi/abs/10.1504/IJTMCC.2013.052524); R. Gajanayake; R. Iannella; T. Sahama («Sharing With Care: An Information Accountability Perspective», IEEE Internet Computing, vol. 15, 4, 2011.)

[21] Cf. Observacom ("Ott Regulation. Key Points For The Democratic Regulation Of «over-the-top» Services So As To Ensure A Free And Open Internet And The Full Exercise Of Digital Rights And Freedom Of Expression", cit.), 9 ("Such intermediaries no longer provide just technical support and 'transit routes', but often affect the contents that circulate through such routes. Not only are they able to monitor all content produced by third parties, but they can also intervene in them, ordering and prioritizing their access and, therefore, determining what contents and sources of information a user may or may not view. They can also block, eliminate or de-index content—such as speeches protected by the right to freedom of expression—as well as users' accounts or profiles. These actions are often forced by external pressures from government authorities or other private actors, but also by the decisions taken by the intermediaries themselves"). *See also* L. Lessig (*Code*, cit.), Chap. 6 (on the history of content moderation in the first Internet "communities") and S.T. Roberts ("Content Moderation", 2017.), 3-4 (calling attention to a "new scale" of moderation as a consequence of concentration, driven in part by the "legal, financial and reputational" risks that these public companies that respond to their shareholders must face due to the content generated by users without any control).

[22] Cf. D. Ghosh; B. Scott, «Digital Deceit: The Technologies Behind Precision Propaganda On The Internet». New America, Washington D.C. January 2018.

[23] This model had already been pointed out by Lessig. Cf. L. Lessig, *Code*, cit., Chap. 11.

have published studies and provided information as part of transparency efforts aimed at placating their critics. They supported, and in some cases led, programs to strengthen quality journalism and "fact-checking" as less-problematic remedies.

These reactions cannot be explained outside of the pessimistic turn previously described. What is being questioned is not only the role they play in the flow of information online but also the regulatory framework that, until now, determined the obligations and responsibilities of intermediaries on the Internet. The principle of non-liability of intermediaries made sense in a decentralized network such as the one represented in Figure 1: establishing responsibilities in this scenario implied hindering, rather than favoring, the free flow of information. But in a centralized network, does it make sense to keep the same principle?

The role that companies assume in matters of disinformation — in many cases, under increasing social and political pressure — poses a more general problem. This problem can be evidenced by the following question: who dealt with disinformation before the Internet? The freedom of expression standards denied any authority to the state to define what is true and what is false: that judgment was reserved for each of the citizens who, within the framework of the free market of ideas, could access different points of view and form their own opinion. This arrangement was not only a sign of confidence in a virtuous citizenship: it is a rule for the adjudication of powers and functions in a democratic society. The state should not intervene in determining what is true and what is false, and that explains why the jurisprudence on freedom of expression of the 20th century was so protective of this right around the world.

The platforms' current responses to disinformation aggravate this state of affairs for two reasons: first, because these companies are in part driven by state demands for greater action, which violates the principle detailed above, regarding the role of the state in the discernment between what is true and what is false. Suddenly, the state has informally legitimized — through demands and pressure from some of its officials — an intermediary actor, central to the circulation of information, as the arbiter of the truth. In the Internet age, it seems that citizens can no longer do what they did before. Second, as private companies, they carry-out these actions within the framework of a legally protected right of contractual roots, without adequate mechanisms of accountability or liability. In other words, we have added a new actor who makes decisions that used to be up to each individual, and it is not even an actor we can influence through democratic procedures.

Therefore, these challenges present a problem to be solved: do we maintain our commitment to broad freedom of expression and trust again that the best answer to false information is free competition from the marketplace of ideas, or do we challenge this view? In the second case, what do we replace it with?

## B. Actions

To document the actions deployed by the platforms in recent years we divided them into four categories:

1. *Awareness actions* Here we group the platforms' political actions, including partnerships with other actors, the promotion of education campaigns, digital and media literacy, and so on. These are actions that seek to build a positive ecosystem regarding disinformation, to empower actors and strategies that are expected to combat the phenomenon. By definition, they do not involve changes to the platforms' code or policies.
2. *Changes to the code.* These are actions that modify the code of the platforms, that is, the way they work. This includes changes to the algorithms for recommendations, visibility, and scope of content.
3. *Policy changes and moderation actions.* Here we refer to the actions that modify the policies of the companies, both the Terms of Service (TOS) and the content moderation policies (community guidelines). In this category, we also group actions that implement internal policies or external mandates for the removal of content reported as illegal or contrary to community guidelines, in cases of disinformation.
4. *Transparency and public relations actions.* This group includes the actions that reveal information about the operation of the platform, both those generated directly by the platforms and those that are the result of independent researchers with access to privileged information. This category also groups abstract or wishful statements from platforms on how to deal with the challenge of disinformation.

These companies did not deploy all the actions at the same time: each action has its history, linked to the temporary nature of the phenomenon of disinformation and the changes in attitude towards the Internet on the optimism/pessimism

binary. Their objective was to face the demands of users, civil society, and governments concerning the central role they acquired in the circulation of information on the Internet as a consequence of their success.

Awareness actions (1) operate at a level that involves the companies' investment of resources but do not impact in any way their business model. The same occurs, although to a lesser extent, with transparency and public relations actions (4), which can only be viewed as "costly" from a free competition point of view (they reveal information to adversaries or competitors). Moderation actions involve the design and implementation of content-related policies, such as community guidelines or terms of service (TOS), which can entail high costs depending on the complexity of the structure (3). Finally, changes in the code affect the users' operation of the services offered by companies and, while these modifications can have an unexpected disruptive effect, they should be seen as the most "radical" actions within the menu of options available (2).[24]

## *1. Awareness actions*

These are actions aimed at alerting users about disinformation campaigns or promoting quality journalism as a presumably effective answer to the problem. These actions also bring together the various partnerships established by the companies with other actors, especially with traditional communication media and verification organizations.

### *a. Support and promotion of "quality" journalism*

Google and Facebook have supported initiatives to strengthen "quality" journalism. Both companies focused those efforts on two initiatives that seem quite

[24] Mark Zuckerberg's announcement on November 19, 2016, about ten days after the presidential election that brought Donald Trump to the presidency, is quite revealing of what came next. These were the measures: (1) stronger detection, through automated mechanisms; (2) ease of reporting, to facilitate "detection;" (3) notices - Zuckerberg said he was "exploring" the possibility of labeling stories that had been "reported" as false by "third parties or our community;" (4) partnerships with "fact-checkers" and — in general — listening more to journalism and the verification industry; (5) recommendation of "quality" articles (2). Zuckerberg announced that he was going to "raise the bar" for the quality of articles featured in "related articles;" (6) changes in ad policies to "alter" the "fake news economy" (3). *See* M. Zuckerberg (Mark Zuckerberg On Disinformation, Facebook status update, 11/19/2016, retrieved from https://www.facebook.com/zuck/posts/10103269806149061 Last access: March/2/2020.).

similar: the Facebook Journalism Project[25] and the Google News Initiative.[26]

In general, most of these actions took place in the United States, Canada, and the United Kingdom, but they have also spread to other parts of the world. In October 2018, for example, the Facebook Journalism Project announced support for Arab journalists to fight disinformation[27] and two years later it announced the expansion of its "accelerator" program globally.[28] In Latin America, they seem to have reached some Brazilian media outlets.[29] *The Google News* Initiative was behind the launch of *Reverso* in Argentina[30] and supported *Impacto.jor*, a project aimed at measuring the impact of the media.

This type of action is based on this premise: traditional media outlets can be important players in fighting misinformation. However, this is not entirely evident. For example, we need to analyze particular situations to understand whether or not and to what extent these media outlets retain their role as gatekeepers. On the other hand, it is also not clear that traditional media always operate in opposition to disinformation campaigns: for example, the study by Benkler, Faris, and Roberts on the circulation of information in the United States revealed that a traditional news network such as Fox News played a central role in leveraging

---

[25] See, e.g., Brown Campbell (Campbell Brown: Facebook Is Doing More To Support To Local News, Campbell Brown: Facebook Is Doing More to Support to Local News, 01/15/2019, retrieved from https://www.facebook.com/journalismproject/facebook-supports-local-news Last access: March/11/2020.); J. Mabry; D. Mendoza (Facebook Develops Training And Support For Local Newsrooms, Facebook Develops Training and Support for Local Newsrooms, 04/27/2017, retrieved from https://www.facebook.com/journalismproject/facebook-training-support-local-newsrooms Last access: March/2/2020.); F. Dinkelspiel (*Berkeleyside To Launch News Site In Oakland, Will Convert To Nonprofit*, Berkeleyside, 12/10/2019, retrieved from https://www.berkeleyside.com/2019/12/10/berkeleyside-to-launch-news-site-in-oakland-will-convert-to-nonprofit Last access: August/27/2020.). Facebook and Mozilla also supported the News Integrity Initiative at CUNY's journalism school. *See* Poynter (Can Trust In The News Be Repaired? Facebook, Craig Newmark, Mozilla And Others Are Spending $14 Million To Try, Poynter, 04/03/2017, retrieved from https://www.poynter.org/tech-tools/2017/can-trust-in-the-news-be-repaired-facebook-craig-newmark-mozilla-and-others-are-spending-14-million-to-try/ Last access: March/2/2020.); J. Rutenberg; M. Isaac («Facebook Hires Campbell Brown To Lead News Partnerships Team», The New York Times, 1/6/2017, retrieved from https://www.nytimes.com/2017/01/06/business/media/facebook-campbell-brown-media-fake-news.html Last access: March/02/2020.)

[26] *See* https://newsinitiative.withgoogle.com/

[27] MENA Herald, New Program Will Help Arab Journalists Sort Fact From Fiction On Social Media, MENA Herald, 11/28/2018, retrieved from https://www.menaherald.com/en/business/media-marketing/new-program-will-help-arab-journalists-sort-fact-fiction-social-media Last access: March/11/2020.).

[28] D. Grant, What's Next For Facebook's Accelerator Program, Facebook Accelerator Program: What's Planned for 2020, 01/28/2020, retrieved from https://www.facebook.com/journalismproject/programs/accelerator/whats-next-2020 Last access: March/13/2020.).

[29] Cf. *Ibid.*

[30] P. J. Blanco, «*La Elección De Las "fake News": Más De 100 Medios Lanzan "reverso", Para Combatir La Desinformación En Campaña*» [The Elections Of "fake News:" More Than 100 Media Outlets Launch "Reverso," To Combat Disinformation In Campaign], Diario Clarín, 6/11/2019, retrieved from https://www.clarin.com/politica/eleccion-fake-news-100-medios-lanzan-reverso-grupo-combatir-desinformacion-campana_0_QYwSO8kRP.html Last access: March/13/2020.).

President Trump's disinformation strategy in the face of the Robert Muller investigation.[31]

### *b. Raise user awareness*

Other actions of the platforms sought to make users aware of the existence of disinformation itself. This type of action rests on the paradigm described in the previous section, which set forth that it is up to citizens to define what is true and what is false. The objective seems twofold: to educate users who are unsuspecting or surprised by the chaotic circulation of information that seems to increase every day, and to remind citizens of a democratic political community that they have a duty to critically address public debate.

Thus, in April 2017 Facebook, in collaboration with First Draft, released in fourteen countries a series of "tips" to identify fake news.[32] The same month, it launched campaigns in German newspapers to raise awareness about the phenomenon.[33] YouTube also announced workshops with teenagers in the UK[34] and Google launched a similar global initiative.[35] Twitter also promoted a *Media Lit-*

---

[31] Y. Benkler; R. Faris; H. Roberts, Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics, Oxford University Press, New York, NY, 2018, Page 158 («…vemos a Fox News tomando un rol central de una forma clara y distintiva, desviando la atención y la culpa, interpretando la investigación en términos profundamente partidarios, y sembrando confusión dudas. Y es aquí también dónde el marco amplio y ambiguo adquiere caracteres de falsificación, cuyos patrones encajan con los patrones de campañas sostenidas de desinformación por actores de propaganda, más que con errores episódicos de un actor periodístico…» [we see Fox News taking center stage in a much clearer and more distinctive way by deflecting attention and blame, interpreting the investigation in deeply partisan terms, and sowing confusion and doubt. And it is here too that the broad and loose gestalt frame takes on distinct falsifiable forms whose pattern fits that of a sustained disinformation campaign by a propaganda outlet, rather than as episodic errors by a journalistic outlet...").

[32] A. Mosseri, A New Educational Tool Against Misinformation, About Facebook, 04/06/2017, retrieved from https://about.fb.com/news/2017/04/a-new-educational-tool-against-misinformation/ Last access: March/2/2020 («Improving news literacy is a global priority, and we need to do our part to help people understand how to make decisions about which sources to trust»); C. Silverman, Facebook Wants To Teach You How To Spot Fake News On Facebook, BuzzFeed News, 04/06/2017, retrieved from https://www.buzzfeednews.com/article/craigsilverman/facebook-wants-to-teach-you-how-to-spot-fake-news-on#.erZ2mjEY2a Last access: March/2/2020.).

[33] «Facebook Buys Full-page Ads In Germany In Battle With Fake News», Bloomberg.com, 4/13/2017, retrieved from https://www.bloomberg.com/news/articles/2017-04-13/facebook-buys-full-page-ads-in-germany-in-battle-with-fake-news Last access: March/2/2020.).

[34] BBC, Youtube To Offer Fake News Workshops To Teenagers, BBC Newsbeat, 04/20/2017, retrieved from http://www.bbc.co.uk/newsbeat/article/39653429/youtube-to-offer-fake-news-workshops-to-teenagers Last access: March/2/2020 («Naomi Gummer, head of public policy at YouTube UK, said: "The internet is what we want it to be. It can be an unpleasant place where people misunderstand and deliberately deceive each other. Or it can be this amazing place where we can share, collaborate, understand and help each other").

[35] P. Diwanji, «be Internet Awesome»: Helping Kids Make Smart Decisions Online, Google, 06/06/2017, retrieved from https://blog.google/technology/families/be-internet-awesome-helping-kids-make-smart-decisions-online/ Last access: March/4/2020.).

*eracy Week* in some of its markets.[36]

These types of actions are not very problematic but the evidence regarding their effectiveness is contradictory. Efforts to counter false information have been shown to sometimes backfire.[37] However, preventive intervention seems to have positive effects in some cases, as it puts people on the defensive from a cognitive perspective, and it could be effective in preventing possible epistemic effects of disinformation campaigns.[38]

### *c. Partnerships with verifiers and other actors*

Partnering with organizations that verify information seems to have risen as the platforms' preferred response to face the problem of disinformation. Facebook announced partnerships with the *Associated Press*,[39] with *Boom* to work in the Indian market,[40] with the IFCN to improve news checking, [41] and with the Reuters agency.[42] In October 2018 Google launched a search engine specialized in verifiers, including Latin American actors.[43] There seems to be a close working relationship between platforms and verification agencies, which proliferated in recent times, as evidenced by *Reverso* project in Argentina.

---

[36] Twitter public policy (Update: Russian Interference In The 2016 Us Presidential Election, Twitter Blog, 09/28/2017, retrieved from https://blog.twitter.com/en_us/topics/company/2017/Update-Russian-Interference-in-2016--Election-Bots-and-Misinformation.html Last access: March/4/2020.); *see* https://medialiteracyweek.us/wp-content/uploads/2015/07/2017-media-literacy-week-press-release-final.pdf.

[37] D. J. Flynn; B. Nyhan; J. Reifler, «The Nature And Origins Of Misperceptions: Understanding False And Unsupported Beliefs About Politics», Political Psychology, vol. 38, S1, 2017, retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/pops.12394; *See* B. Nyhan; J. Reifler, «Estimating Fact-checking's Effects». American Press Institute, Arlington, V.A. 2015.

[38] Cf. J. Cook; S. Lewandowsky, «Misinformation And How To Correct It», in Robert A Scott, Stephan M Kosslyn (eds.) Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource, 1, Wiley, 2015, pages 6-7 retrieved from https://onlinelibrary.wiley.com/doi/book/10.1002/9781118900772. In addition to this strategy, Cook and Lewandowski point out that the force with which false information is refuted and the alternative explanation that "fills the gap" left by the information withdrawn from the mental models used to understand the information in question were effective, under experimental conditions, to strengthen the rebuttal effect.

[39] AP, Ap To Debunk Election Misinformation On Facebook, Associated Press, 03/07/2018, retrieved from https://www.ap.org/press-releases/2018/ap-to-debunk-election-misinformation-on-facebook Last access: March/9/2020.).

[40] P. Dixit, Facebook Has Finally Started Fact-checking Fake News In Its Largest Market, BuzzFeed News, 4/17/2018, retrieved from https://www.buzzfeednews.com/article/pranavdixit/facebook-has-finally-started-fact-checking-fake-news-in-its Last access: March/9/2020.).

[41] Poynter, Ifcn And The Facebook Journalism Project Announce Fact-checking Innovation Initiative, Poynter, 11/05/2019, retrieved from https://www.poynter.org/fact-checking/2019/innovation-initiative/ Last access: August/ 27/2020.).

[42] D. Patadia, «Reuters Launches Fact-checking Initiative To Identify Misinformation, In Partnership With Facebook», Reuters, 2/12/2020, retrieved from https://www.reuters.com/article/rpb-fbfactchecking-idUSKBN2061TG Last access: August/27/2020.).

[43] Retrieved from: https://toolbox.google.com/factcheck/

These types of partnerships also do not present significant problems from the point of view of freedom of expression, and positive effects have been reported.[44] However, many challenges persist, including its limited impact on the less informed or less educated, who — in turn — are more likely to consume false information.[45]

## 2. Code related actions

This group identifies the actions that modify in some way the *code* of the platform. Detection, reporting, and recommendations for additional information are typical actions in this category.

### *d. Detection of false information*

The horizontal information production model transforms large platforms into channels of content created by third parties. But, as noted above, they increasingly play an active curating and moderating role (Table 1). Companies have shaped their moderation policies for their own economic interests and due to pressure from external actors, like the legal requirements in various countries to combat forms of speech clearly banned such as child pornography or threats; the social pressure to combat problematic speech, although not necessarily illegal, such as the so-called discriminatory discourse; and the political pressure to tackle, for example, problems like the one this study analyzes.

**Table 1 - Reasons that promote content moderation by intermediaries.**

| Reasons | Positive | Negative |
|---|---|---|
| **Business** | the introduction of algorithms as a way to generate more relevant content and increase user engagement | when companies prefer that certain forms of speech not be on their platforms |
| **Social** | when society asks for certain forms of speech (e.g. public good awareness campaigns) | when society rejects certain forms of speech and calls for action against them (e.g. hate speech) |
| **Political** | those that originate in dialogue and alliances (e.g. electoral authorities) | those that come from pressure from the political system (legislative investigations) |
| **Legal** | the demands made by certain content or practices | those that establish obligations in terms of prohibited speech |

[44] B. Nyhan; J. Reifler, "Estimating Fact-checking's Effects", cit.

[45] Cf. G. Pennycook; D. G. Rand, «Who Falls For Fake News? The Roles Of Analytic Thinking, Motivated Reasoning, Political Ideology, And Bullshit Receptivity», SSRN Electronic Journal, 2017, retrieved from https://www.ssrn.com/abstract=3023545.

Within the framework of these general moderation actions, some actions aim at "detecting" false information. These are actions that resort to different methods, but they start from the same premise: disinformation is a type of undesirable speech within these services.[46] In the implementation of this policy, the first challenge was always to identify false content with some degree of precision, and for this, the platforms seem to have resorted to two strategies: user reporting and automatic detection.

The first approach simply asks users to report content that is suspected of being false, e.g., through "surveys" in which feedback is requested[47] or giving the possibility of "reporting" content as false. Google implemented the option to report automatic recommendations when entering text in the search engine[48] and Twitter set up — in April 2019 — a new reporting modality that includes "false information regarding elections,"[49] although this option is not yet available in Latin America.[50]

The second approach is based on the development of technological tools capable of automatically detecting false content, based on algorithms, artificial intelligence, and machine learning. Thus, for example, Facebook announced the adoption of cri-

---

[46] This moment appears to be linked to the 2016 US presidential election. *See* T. Gillespie, Custodians Of The Internet: Platforms, Content Moderation, And The Hidden Decisions That Shape Social Media, Yale University Press, New Haven, 2018, pp. 65-66. This point is interesting because there was some ambiguity on the part of the platforms regarding their own capabilities to "monitor" that content.

[47] *See* C. Zang; S. Chen, Using Qualitative Feedback To Show Relevant Stories, About Facebook, 02/01/2016, retrieved from https://about.fb.com/news/2016/02/news-feed-fyi-using-qualitative-feedback-to-show-relevant-stories/ Last access: March/2/2020.).

[48] Google's «project Owl» -- A Three-pronged Attack On Fake News & Problematic Content, Search Engine Land, 04/25/2017, retrieved from https://searchengineland.com/googles-project-owl-attack-fake-news-273700 Last access: May/7/ 2020.

[49] Twitter Safety, Strengthening Our Approach To Deliberate Attempts To Mislead Voters, Twitter blog, 04/24/2019, retrieved from https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-approach-to-deliberate-attempts-to-mislead-vot.html Last access: March/11/2020.).

[50] Hugo TwitterGov, *Calidad de la información durante elecciones [*Quality of information during elections*]*, 2020 (on file with the author), pp.4-5 ("*Es importante que esta política sea aplicada de manera global y si bien la herramienta como tal no está* disponible *todavía en Latinoamérica, el plan es implementarla en todo el mundo tan pronto nos sea posible. Debemos mencionar no obstante que la ausencia de la herramienta de reporte público hasta el momento, esto en sentido alguno implica la ausencia de aplicación de esta política, en tanto: i) Twitter aplica sus reglas de manera proactiva (más del 50% de los contenidos accionados por nosotros son detectados de manera proactiva, sin mediar reporte de algún usuario) y ii) el equipo de Política Pública para Latinoamérica Hispanoparlante trabaja constantemente para establecer un canal de comunicación directo con las diferentes autoridades electorales de la región (por ejemplo Argentina, Colombia y México) para reportar contenido violatorio de nuestras políticas de contenido publicitario o de las regulaciones locales en este sentido*" [It is important that this policy be applied globally and although the tool as such is not yet available in Latin America, the plan is to implement it all over the world as soon as possible. However, we must mention that the absence of the public reporting tool so far in no sense implies the absence of implementation of this policy, insofar as: i) Twitter applies its rules proactively (more than 50% of the content questioned by us is detected proactively, without any user reporting) and ii) the Public Policy team for Spanish-speaking Latin America constantly works to establish a channel of direct communication with the different electoral authorities in the region (for example Argentina, Colombia and Mexico) to report content that violates our policies regarding advertising content or local regulations in this regard]).

teria based on user behavior to classify content as sensationalist or non-sensationalist[51] (although it changed its approach after a satirical publication was incorrectly classified as false[52]). On the other hand, in January 2019, YouTube announced that it would seek to "turn down the volume" of content that, although it does not violate its internal policies, is "close" to violating them.[53] The commitment to this type of tool is based on the potential scalability of a successful solution. However, it is difficult: detecting fake content is complex and even when it is relatively easy (for example, in the case of deepfakes) it is very likely that it requires human intervention to, among other things, distinguish merely satirical content from a disinformation campaign.[54]

### *e. Giving context to information by "labeling" content*

Another action that various platforms have adopted in recent times has to do with the "labeling" of content to give more context to the information. Thus, for example, in June 2019 Twitter announced that it would label posts by politicians or public figures that violate their rules and that, in other circumstances, they would be deplatformed.[55] This action became especially controversial when it began labeling posts by then-United States President Donald Trump.[56] Facebook

---

[51] M. Zuckerberg, Building Global Community, Facebook, 02/16/2017, retrieved from https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634 Last access: March/2/2020.).

[52] E. Wemple, «Facebook Working On Approach To Classifying Satirical News Pieces - The Washington Post», The Washington Post, 3/5/2018, retrieved from https://www.washingtonpost.com/blogs/erik-wemple/wp/2018/03/05/facebook-working-on-approach-to-classifying-satirical-news-pieces/ Last access: May/7/ 2020.

[53] YouTube, Continuing Our Work To Improve Recommendations On Youtube, Official YouTube Blog, 01/25/2019, retrieved from https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html Last access: March/11/2020.).

[54] *See*, e.g., a Facebook progress report in this regard in T. Lyons (Increasing Our Efforts To Fight False News, About Facebook, 06/21/2018, retrieved from https://about.fb.com/news/2018/06/increasing-our-efforts-to-fight-false-news/ Last access: July/14/2020.). On the challenges of AI to fight misinformation, *see* S. C. Woolley (We're Fighting Fake News Ai Bots By Using More Ai. That's A Mistake., MIT Technology Review, 01/08/2020, retrieved from https://www.technologyreview.com/2020/01/08/130983/were-fighting-fake-news-ai-bots-by-using-more-ai-thats-a-mistake/ Last access: July/14/2020.).

[55] E. Dwoskin; T. Romm «Twitter Adds Labels For Tweets That Break Its Rules — A Move With Potentially Stark Implications For Trump's Account», The Washington Post, 06/27/2019, retrieved from https://www.washingtonpost.com/technology/2019/06/27/twitter-adds-labels-tweets-that-break-its-rules-putting-president-trump-companys-crosshairs/ Last access: March/11/2020.).

[56] D. Alba; K. Conger; R. Zhong, «Twitter Adds Warnings To Trump And White House Tweets, Fueling Tensions», The New York Times, 05/29/2020, retrieved from https://www.nytimes.com/2020/05/29/technology/trump-twitter-minneapolis-george-floyd.html Last access: July/14/2020; J. Byrnes, Twitter Flags Trump Tweet On Protesters For Including «threat Of Harm», TheHill, 06/23/2020, retrieved from https://thehill.com/policy/technology/504133-twitter-flags-trump-tweet-on-protesters-for-including-threat-of-harm Last access: July/14/2020; K. Conger, «Twitter Had Been Drawing A Line For Months When Trump Crossed It», The New York Times, 05/30/2020, retrieved from https://www.nytimes.com/2020/05/30/technology/twitter-trump-dorsey.html Last access: July/14/2020; A. Wiener, Trump, Twitter, Facebook, And The Future Of Online Speech, The New Yorker, 07/06/2020, retrieved from https://www.newyorker.com/news/letter-from-silicon-valley/trump-twitter-facebook-and-the-future-of-online-speech Last access: July/10/2020.

announced in June 2020 that it would label media outlets belonging to foreign states such as Russia's Sputnik agency [57]and in August Twitter made a similar announcement.[58] Likewise, platforms also announced that they would identify the points of origin of various pieces of content.[59] Furthermore, Twitter announced that in certain cases it would ask users to confirm their intention to re-post certain information before uploading it.[60]

Giving context to information reached new levels during the 2020 US presidential election when it began to label posts by former President Donald J. Trump that deliberately misinformed about alleged "fraud" in the elections[61].

This last case highlights some of the problems involved in these actions aimed at demanding context. If it is limited to providing more context, it is not very risky from the point of view of freedom of expression, but it could be problematic if that labeling has any effect on the way information circulates; for example, if it negatively affects the content recommendation algorithm. On the other hand, these moderation actions face a considerable challenge in terms of scale, replicability, and management of the standards used to identify each case: the consistent application of the norm in different countries and different contexts promises to be a considerable challenge and numerous critics have pointed out that while Trump's tweets deserved labeling, similar tweets from other political or opinion leaders did not get the same treatment.

---

[57] N. Gleicher, Labeling State-controlled Media On Facebook, About Facebook, 06/04/2020, retrieved from https://about.fb.com/news/2020/06/labeling-state-controlled-media/ Last access: August/ 27/2020.

[58] Twitter Support, New Labels For Government And State-affiliated Media Accounts, Twitter Blog, 08/06/2020, retrieved from https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts.html Last access: August/ 27/2020.

[59] A. Joseph, Making Pages And Accounts More Transparent, About Facebook, 04/22/2020, retrieved from https://about.fb.com/news/2020/04/page-and-account-transparency/ Last access: August/27/2020; R. Wong, Google's Taking Another Big Step To Stop The Spread Of Fake News, Mashable, 12/17/2017, retrieved from https://mashable.com/2017/12/17/google-news-no-hiding-country-origin-stop-fake-news/ Last access: March/6/2020.

[60] Twitter Support, Twitter Support On Twitter: "Sharing An Article Can Spark Conversation, So You May Want To Read It Before You Tweet It. To Help Promote Informed Discussion, We're Testing A New Prompt On Android — When You Retweet An Article That You Haven't Opened On Twitter, We May Ask If You'd Like To Open It First., Twitter, 06/10/2020, retrieved from https://twitter.com/TwitterSupport/status/1270783537667551233 Last access: August/ 27/2020.

[61] The situation persistently escalated, with labels as the preferred tool, until, on January 6, 2021, a demonstration before the United States Congress advanced on the security forces and broke into the compound where the election results were being validated. After the incident, various platforms suspended the president's account on the grounds that he had incited this "insurrection" in part through his instances of disinformation. In the case of Facebook, its Oversight Board decided to take the case to rule on the matter. See S. Tavernise; M. Rosenberg, "These Are the Rioters Who Stormed the Nation's Capitol," The New York Times, 1/8/2021, retrieved from https://www.nytimes.com/2021/01/07/us/names-of -rioters-capitol.html Last access: February 2, 2021; D. Ghosh; J. Hendrix, Facebook's Oversight Board Takes on the Curious Case of Donald J. Trump, Verfassungsblog, 01/29/2021, retrieved from https://verfassungsblog.de/fob-trump/ Last access: February/2/2021.

### *f. More quality journalism and verification*

Sometimes, partnerships with verifiers involve changes in the code of the platforms. For example, Google announced that verifiers would receive a privileged place in search results when the referred information has been duly checked by a recognized verifier,[62] a change that — later on — it also introduced in Google News.[63] It also announced that it would give priority to quality journalism in its search engine,[64] something that Facebook also announced concerning changes in its algorithm.[65]

In April 2017, Facebook announced that "related articles" would have a more prominent position, before users read certain content.[66] This would include the active promotion of verified information when automated detection mechanisms send suspicious content to independent verifiers and — after the corresponding "check" — those articles could be displayed below the original article.[67] Likewise, it adopted actions against cloaking, a technique used to avoid internal controls,[68] *spamming*,[69] and *clickbait*.[70] In October 2017 Facebook also announced a new button, in the testing stage, to offer more contextual information on shared

[62] Cf. A. Mantzarlis (Google Is Now Highlighting Fact Checks In Search, Poynter, 04/07/2017, retrieved from https://www.poynter.org/fact-checking/2017/google-is-now-highlighting-fact-checks-in-search/ Last access: March/2/2020.); A. Mantzarlis (*Google News Now Has A «fact Check» Tag*, Poynter, 10/13/2016, retrieved from https://www.poynter.org/fact-checking/2016/google-news-now-has-a-fact-check-tag/ Last access: March/2/2020.). The behavior could not be verified, it depends on the use of various technologies such as e.g. schema.org

[63] J. Lichterman, Google News Launches A Streamlined Redesign That Gives More Prominence To Fact Checking, Nieman Lab, 06/27/2017, retrieved from https://www.niemanlab.org/2017/06/google-news-launches-a-streamlined-redesign-that-gives-more-prominence-to-fact-checking/ Last access: March/4/2020.

[64] R. Gingras, Elevating Original Reporting In Search, Google blog, 09/12/2019, retrieved from https://blog.google/products/search/original-reporting Last access: March/11/2020; D. Osborn, Smarter Organization Of Top Stories In Search, Google, 12/11/2019, retrieved from https://blog.google/products/search/smarter-organization-top-stories-search/ Last access: August/ 27/2020.

[65] S. Fischer, Axios Media Trends: Facebook To Change Its Algorithm, Axios, 06/30/2020, retrieved from https://www.axios.com/newsletters/axios-media-trends-40d2f971-c434-4359-8a0f-7fa4155626bc.html Last access: August/27/2020.

[66] S. Su, New Test With Related Articles, About Facebook, 04/25/2017, retrieved from https://about.fb.com/news/2017/04/news-feed-fyi-new-test-with-related-articles/ Last access: March/2/2020.

[67] *Ibid.* («Now, we will start using updated machine learning to detect more potential hoaxes to send to third-party fact checkers. If an article has been reviewed by fact checkers, we may show the fact checking stories below the original post. In addition to seeing which stories are disputed by third-party fact checkers, people want more context to make informed decisions about what they read and share. We will continue testing updates to Related Articles and other ongoing News Feed efforts to show less false news on Facebook and provide people context if they see false news»).

[68] R. Leathern, Addressing Cloaking So People See More Authentic Posts, About Facebook, 08/09/2017, retrieved from https://about.fb.com/news/2017/08/news-feed-fyi-addressing-cloaking-so-people-see-more-authentic-posts/ Last access: March/4/2020.

[69] A. Mosseri, Showing More Informative Links In News Feed, About Facebook, 06/30/2017, retrieved from https://about.fb.com/news/2017/06/news-feed-fyi-showing-more-informative-links-in-news-feed/ Last access: March/4/2020.

[70] A. Babu; A. Liu; J. Zhang, New Updates To Reduce Clickbait Headlines, About Facebook, 05/17/2017, retrieved from https://about.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/ Last access: March/4/2020.

articles,[71] which later became the Trust Indicators.[72] Google also developed a similar tool (Knowledge Panels) that allows its search engine to learn more about the origin of the information.[73] In March 2018, YouTube announced that Wikipedia would link information to videos that promoted conspiracy theories[74] and in February 2019, it made a similar announcement regarding India, linked to the publication of fact-checking.[75]

### *g. Banning some behaviors*

One of the platforms' preferred strategies to combat disinformation is to target user "behaviors" that may be related to disinformation campaigns. Thus, for example, it is possible that these types of campaigns are powered by anonymous accounts, that leave traces of specific behaviors - an unreasonable tendency to retweet, to send or spread information, to seek to install trends -, or that they are simply bots, that is, accounts driven by a code, automatically (for example, retweeting some accounts or content or publishing periodic information of certain characteristics, etc.). This strategy is to a certain extent "preferred" by platforms because it avoids the thorny issue of defining disinformation: it is no longer necessary to determine whether certain content is false, but it is enough to detect prohibited behaviors, a task that can be done on an objective basis and through automated processes.

Twitter made this particular point.[76] Among its main strategies to combat misinformation, it included not only the detection of behaviors that reveal practices that violate the company's TOS but also attempts to prevent malicious actors from "beating"

---

[71] A. Anker, New Test To Provide Context About Articles, About Facebook, 10/05/2017, retrieved from https://about.fb.com/news/2017/10/news-feed-fyi-new-test-to-provide-context-about-articles/ Last access: March/6/2020.

[72] A. Anker, Launching New Trust Indicators From The Trust Project For News On Facebook, Facebook For Media, 11/17/2017, retrieved from https://www.facebook.com/facebookmedia/blog/launching-new-trust-indicators-from-the-trust-project-for-news-on-facebook Last access: March/6/2020 («We are initially testing these Trust Indicators with a small group of publishers, with plans to expand more broadly over the coming months»).

[73] Google, Learn More About Publishers On Google, Google, 11/17/2017, retrieved from https://blog.google/products/search/learn-more-about-publishers-google/ Last access: March/13/2020.

[74] C. Newton, Youtube Will Add Information From Wikipedia To Videos About Conspiracies, The Verge, 03/13/2018, retrieved from https://www.theverge.com/2018/3/13/17117344/youtube-information-cues-conspiracy-theories-susan-wojcicki-sxsw Last access: March/9/2020.

[75] Dixit, Pranav, Youtube Is Rolling Out A Feature That Shows Fact Checks When People Search For Sensitive Topics, Buzzfeed News, 03/07/2019, retrieved from https://www.buzzfeednews.com/article/pranavdixit/youtube-debunk-information-panels-india Last access: March/11/2020.

[76] Twitter public policy, *Update*, cit.

the system e.g., manipulating the trending topics.[77] On this issue, Twitter reported:

> "We interpret "manipulating the platform" as using Twitter for the purpose of artificially amplifying or suppressing information or carrying out actions that manipulate or hinder the users' experience on Twitter. In this regard, we ban massive, high-intensity, or deceptive actions that confuse users or hinder their experience. There are many ways to manipulate the platform, and our rules are intended to counteract a wide variety of prohibited behaviors, for example, spam for commercial purposes, which generally aims to divert traffic or attention from a Twitter conversation to other accounts, websites, products, services or initiatives; fake interactions that aim to make accounts or content appear more popular or active than they actually are; and coordinated activities that aim to artificially influence conversations through the use of multiple accounts, fake accounts, automatic actions or scripts.[78]

Facebook opts for a similar model, but based on a different policy: anonymous accounts are banned there - Facebook wants each of its users to tie their real identity to their account.[79] As in the case of Twitter, the measures taken by Facebook in this regard allow it to avoid having to analyze the content of the posts, leading to (a) an automated moderation tool that (b) does not depend on complicated content analysis.[80] Thus, for example, in August 2017 the company announced

---

[77] N. Confessore; G. J. X. Dance, «Battling Fake Accounts, Twitter To Slash Millions Of Followers», The New York Times, 7/11/2018, retrieved from https://www.nytimes.com/2018/07/11/technology/twitter-fake-followers.html Last access: March/11/2020; Cf. Twitter public policy, Update, cit.

[78] Hugo TwitterGov, "*Calidad de la información durante elecciones*" [Quality of information during elections], cit., p. 2.

[79] S. Shaik, Improvements In Protecting The Integrity Of Activity On Facebook, Facebook, 04/13/2017, retrieved from https://www.facebook.com/notes/facebook-security/improvements-in-protecting-the-integrity-of-activity-on-facebook/10154323366590766 Last access: March/2/2020 («People come to Facebook to make meaningful connections. From the beginning, we've believed that can only be possible if the interactions here are authentic – and if people use the names they're known by. We've found that when people represent themselves on Facebook the same way they do in real life, they act responsibly. Fake accounts don't follow this pattern, and are closely related to the creation and spread of spam. That's why we're so focused on keeping these inauthentic accounts and their activity off our platform»).

[80] *Ibid.* («We've made improvements to recognize these inauthentic accounts more easily by identifying patterns of activity — without assessing the content itself. For example, our systems may detect repeated posting of the same content, or an increase in messages sent. With these changes, we expect we will also reduce the spread of material generated through inauthentic activity, including spam, misinformation, or other deceptive content that is often shared by creators of fake accounts»); J. Weedon; W. Nuland; A. Stamos, «Information Operations And Facebook». Facebook, Palo Alto, California. April 27, 2017. Pages 10 ("Through technical advances, we are increasing our protections against manually created fake accounts and using new analytical techniques, including machine learning, to uncover and disrupt more types of abuse. We build and update technical systems every day to make it easier to respond to reports of abuse, detect and remove spam, identify and eliminate fake accounts, and prevent accounts from being compromised. We've made recent improvements to recognize these inauthentic

that it would block ads from pages that "repeatedly" share false information.[81]

These types of strategies seem to have been adopted to deal with "false amplifiers," the efforts to viralize certain information through various techniques, such as the creation of false accounts, the coordination of certain messages, memes, and so on.[82] In the case of Facebook — which does not allow anonymous accounts — the removal of false accounts becomes a great proxy to confront this type of behavior, which in social media platforms such as Twitter (which protects users' right to anonymity) seems more complicated.[83] Likewise, Facebook also banned "links" in false advertisements and stories.[84]

On Twitter, the fight against bots is justified as they seek to "undermine the core of the service's functionality."[85] In November 2017, Google announced that it would stop allowing sites that do not reveal their origin to appear in Google News.[86] In March 2018, it announced a series of restriction measures on advertisers who were abusing company policies.[87]

As disinformation campaigns are built on specific information dissemination strategies, it is possible to "combat" disinformation by limiting or fighting against these strategies. But this presents two problems. In the first place, the line that separates —for example— "coordinated actions" from simple political campaigns is very blurred.[88] Likewise, it is not a given that platforms can manage this

---

accounts more easily by identifying patterns of activity — without assessing account contents themselves .10 For example, our systems may detect repeated posting of the same content, or aberrations in the volume of content creation. In France, for example, as of April 13, these improvements recently enabled us to take action against over 30, 000 fake accounts").

[81] S. Shukla; T. Lyons, Blocking Ads From Pages That Repeatedly Share False News, About Facebook, 08/28/2017, retrieved from https://about.fb.com/news/2017/08/blocking-ads-from-pages-that-repeatedly-share-false-news/ Last access: March/4/2020 («Today's update helps to disrupt the economic incentives and curb the spread of false news, which is another step towards building a more informed community on Facebook»).

[82] Cf. J. Weedon; W. Nuland; A. Stamos, "Information Operations And Facebook", cit., p. 9.

[83] Cf. *Ibid.* («In the case of Facebook, we have observed that most false amplification in the context of information operations is not driven by automated processes, but by coordinated people who are dedicated to operating inauthentic accounts»).

[84] S. Shukla; T. Lyons, "Blocking Ads From Pages That Repeatedly Share False News", cit.

[85] C. Crowell, Our Approach To Bots And Misinformation, Twitter Blog, 06/14/2017, retrieved from https://blog.twitter.com/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html Last access: March/4/2020.

[86] R. Wong, Google's Taking Another Big Step To Stop The Spread Of Fake News, cit.

[87] I. Lunden (Google Removed 2.3b Bad Ads, Banned Ads On 1.5m Apps + 28m Pages, Plans New Policy Manager This Year, TechCrunch, 03/14/2018, retrieved from http://social.techcrunch.com/2019/03/13/google-removed-2-3b-bad-ads-1-5m-bad-apps-and-28m-bad-pages-plans-new-policy-manager-this-year/ Last access: March/9/2020.); S. Spencer (An Advertising Ecosystem That Works For Everyone, Google - The Keyword, 03/14/2018, retrieved from https://blog.google/technology/ads/advertising-ecosystem-works-everyone/ Last access: 9/March/2020.)

[88] S. McGregor, «What Even Is "Coordinated Inauthentic Behavior" On Platforms?», Wired, 09/17/2020, retrieved from https://www.wired.com/story/what-even-is-coordinated-inauthentic-behavior-on-platforms/ Last access: September/22/2020 («*Esta ambiguedad y el enforcement incosnistente, así como la forma no principista en la que el contenido político es moderado,*

approach easily. On the other hand, it comes at a cost. Consider, for example, the restrictions imposed by WhatsApp on the number of times certain messages can be sent. The restriction assumes that the content being forwarded is problematic, and — by definition — limits the scope of the message. This, in authoritarian countries where encrypted private messaging technology could have some kind of role in the free flow of information, is a problem rather than a solution.

## 3. "Policy" and moderation

The platforms' actions and decisions have to do with their positioning and their principles regarding this issue, or "policy choices." These choices seek to establish a position on the role that they appoint for themselves on the flow of information online and they impact both their internal design and public positions. As we will see below, these positions have changed over time and had cross-cutting impacts, especially in self-regulation and the rules of moderation. Perhaps the most prominent and memorable change to exemplify this point has been the 2019 change in Twitter's slogan: from the guiding principle of speak truth to power, the company moved to "promote a healthy public conversation."[89]

Until 2019, among the analyzed companies, there prevailed a position of refusal to adopt a control role over the veracity or falsehood of information, which resulted in public statements and positions and the adoption of specific rules on the permissible content on their platforms. Thus, for example, Facebook's CEO stated that "we don't check what people say before they say it, and frankly I don't think society wants us to."[90] Shortly after the 2016 presidential election, its CEO

*exacerba las amenazas al proceso electoral, así como la propia capacidad de las plataformas de defenderse de críticos de ambos lados del mostrador*» [This ambiguity and its inconsistent enforcement, as well as the non-principled way in which the political content is moderated, exacerbates the threat to the electoral process, as well as the ability of platforms to defend themselves against critics on both ends of the issue]).

89 CELE, *Comentario en respuesta a los cambios propuestos por Twitter a las reglas sobre contenido deshumanizante* [Commentaries to Twitter's proposed change in rules regarding "Dehumanizing content"], https://www.palermo.edu/Archivos_content/2020/cele/marzo/Internet_y_Derechos_HumanosIII.pdf

90 *"…Zuckerberg además dijo que iba a continuar trabajando con el gobierno de Estados Unidos, iba a profundizar las investigaciones internas, iba a fortalecer el proceso interno de revisión de avisos, duplicar el equipo que trabaja en temas de"integridad de elecciones", expandir alianzas con autoridades electorales y aumentar la cantidad de información sobre amenazas que comparten con otras empresas tecnológicas*. [Zuckerberg also said that he was going to continue working with the United States government, he was going to expand the internal investigations, he was going to strengthen the internal notice review process, double the team that works on issues of "election integrity," expand their partnerships with electoral authorities and increase the amount of information about threats they share with other tech companies]. *See* P. Kafka (Mark Zuckerberg: «freedom Means You Don't Have To Ask For Permission First», Vox, 09/21/2017, retrieved from https://www.vox.com/2017/9/21/16346858/mark-zuckberg-facebook-russia-freedom-permission Last access: 4/March/2020.)

announced general criteria regarding disinformation.[91]

> "The two most discussed concerns this past year were about the diversity of viewpoints we see (filter bubbles) and accuracy of information (fake news). I worry about these and we have studied them extensively, but I also worry there are even more powerful effects we must mitigate around sensationalism and polarization leading to a loss of common understanding. (...) Social media already provides more diverse viewpoints than traditional media ever has. (...) But our goal must be to help people see a more complete picture, not just alternate perspectives. We must be careful how we do this. Research shows that some of the most obvious ideas, like showing people an article from the opposite perspective, actually deepen polarization by framing other perspectives as foreign. A more effective approach is to show a range of perspectives, let people see where their views are on a spectrum and come to a conclusion on what they think is right. Over time, our community will identify which sources provide a complete range of perspectives so that content will naturally surface more."

Specifically on disinformation, Zuckerberg argued that "We are proceeding carefully because there is not always a clear line between hoaxes, satire, and opinion. In a free society, people must have the power to share their opinion, even if others think they are wrong. Our approach will focus less on banning misinformation, and more on emphasizing additional perspectives and information, including that fact-checkers dispute an item's accuracy."[92]

Also in 2017, Facebook launched the Hard Questions Initiative, which sought to ask (and answer) difficult questions about the role of social media in democratic political communities.[93] In October, for example, it revealed information about advertisements linked to Russia during the 2016 election campaign[94] and in November it allowed users to know if they had "liked" any page created by the Inter-

---

[91] M. Zuckerberg, "Building Global Community", cit.

[92] *Ibid.* Along these lines, for example, Facebook announced in April 2017 that it was taking a stronger stance regarding its moderation actions, to include "more subtle forms of misuse, including attempts to manipulate civic discourse and mislead people." J. Weedon; W. Nuland; A. Stamos, "Information Operations And Facebook," cit., p. 3.

[93] E. Schrage, Introducing Hard Questions, About Facebook, 06/15/2017, retrieved from https://about.fb.com/news/2017/06/hard-questions/ Last access: March/4/2020.

[94] E. Schrage, "Hard Questions", cit.

net Research Agency, linked to Russia.[95] In its policy document, Facebook defines "information operations" as "actions taken by organized actors (governmental or non-governmental) to distort domestic or foreign political sentiments, most frequently to achieve a strategic and/or geopolitical outcome. These operations can use a combination of methods, such as fake news, disinformation, or fake account networks aimed at manipulating public opinion (we refer to these as 'false amplifiers.'"[96] Google also produced a short report on "what it found" on its platforms for the same period[97] and Twitter announced that it would follow suit.[98]

These positions may have been based on the optimistic paradigm regarding the Internet that resulted in the non-liability of intermediaries for third-party content. But the pressure on companies during the last three years led them to gradually and reluctantly assume a "controlling" role in matters of disinformation, especially in electoral situations. The changes in this regard are quite significant and show flexible companies with the ability to adapt their policies to an increasing number of demands and play a more active and transparent role in controlling the information that circulates in their platforms.

For example, Facebook seems to have made an effort to defend a particular perspective on freedom of expression that, at least in theory, would continue to justify not having to play this role.[99] However, gradually — and we believe that as a consequence of the increasing pressure on the company — it also adopted changes in its position. For example, in March 2018 Facebook expanded its

---

[95] Facebook, Continuing Transparency On Russian Activity, About Facebook, 11/22/2017, retrieved from https://about.fb.com/news/2017/11/continuing-transparency-on-russian-activity/ Last access: March/6/2020.

[96] *Ibid.*, p. 4. It is interesting to draw attention to Facebook's effort to define precisely the phenomenon, which, for example, distinguishes "information operations" from "fake news" ("News articles that purport to be factual, but which contain intentional misstatements of fact with the intention to arouse passions, attract viewership, or deceive") or« disinformation» (" Inaccurate or manipulated information/content that is spread intentionally. This can include false news, or it can involve more subtle methods, such as false flag operations, feeding inaccurate quotes or stories to innocent intermediaries, or knowingly amplifying biased or misleading information. Disinformation is distinct from misinformation, which is the inadvertent or unintentional spread of inaccurate information without malicious intent").

[97] Google, «What We Found». Google, Palo Alto, California. October 30, 2017.

[98] T. Romm, Twitter Is Exploring How To Notify Its Users That They Saw Russian Propaganda During The 2016 Election, Vox, 01/17/2018, retrieved from https://www.vox.com/2018/1/17/16901428/twitter-notify-users-russia-trolls-2016-election Last access: March/6/2020; Twitter public policy, Update On Twitter's Review Of The 2016 Us Election, 01/19/2018, retrieved from https://blog.twitter.com/en_us/topics/company/2018/2016-election-update.html Last access: March/9/2020.

[99] E. Schrage, Hard Questions: Russian Ads Delivered To Congress, About Facebook, 10/02/2017, retrieved from https://about.fb.com/news/2017/10/hard-questions-russian-ads-delivered-to-congress/ Last access: March/6/2020 («…must also take seriously the crucial place that free political speech occupies around the world in protecting democracy and the rights of those who are in the minority, who are oppressed or who have views that are not held by the majority or those in power»); K. Swisher, Full Transcript: Facebook Ceo Mark Zuckerberg On Recode Decode, Vox, 07/18/2018, retrieved from https://www.vox.com/2018/7/18/17575158/mark-zuckerberg-facebook-interview-full-transcript-kara-swisher Last access: March/11/2020.

content verification policies to photos and videos.[100] If up to that point it was verifying "news," the decision to expand its verification efforts in collaboration with supposedly qualified third parties is a decision that clashes with the more general positioning of the company. In October 2018, it also announced specific efforts to fight disinformation that sought to "suppress" voting,[101] something that it ratified in September 2020 as the elections approached.[102]

In 2019, these policy changes regarding ads during political campaigns became stricter. In January, Facebook announced tougher rules for dealing with advertisements.[103] Google decided to stop broadcasting political advertising in Canada due to transparency rules that would be "very difficult to enforce;"[104] in March 2019 it decided to relaunch its Ad Library[105], and in April 2020 it announced an identity verification system for some of its ads.[106]

Pressure on companies also led to review actions of suspicious activity. In September 2017, Facebook announced that it had identified expenses for more than $ 100,000 from fake accounts from Russia (which it canceled, following its policy of not allowing "inauthentic" accounts).[107] In October 2018, Twitter published a report on "potential information operations."[108] And Facebook gave information on some cases of fake news, reporting what it detected, how it verified the information, and — perhaps more importantly — how it detected it.[109] This develop-

[100] G. Rosen, Hard Questions: What Is Facebook Doing To Protect Election Security?, About Facebook, 03/29/2018, retrieved from https://about.fb.com/news/2018/03/hard-questions-election-security/ Last access: March/9/2020.

[101] J. Menn, «Exclusive: Facebook To Ban Misinformation On Voting In Upcoming U.S. Elections», Reuters, 10/16/2018, retrieved from https://www.reuters.com/article/us-facebook-election-exclusive-idUSKCN1MP2G9 Last access: March/11/2020.

[102] About Facebook, New Steps To Protect The Us Elections, About Facebook, 09/03/2020, retrieved from https://about.fb.com/news/2020/09/additional-steps-to-protect-the-us-elections/ Last access: September/3/2020.

[103] Youtube Just Demonetized Anti-vax Channels, BuzzFeed News, 02/22/2019, retrieved from https://www.buzzfeednews.com/article/carolineodonovan/youtube-just-demonetized-anti-vax-channels Last access: March/11/2020.

[104] T. Cardoso, «Google To Ban Political Ads Ahead Of Federal Election, Citing New Transparency Rules», 3/4/2019, retrieved from https://www.theglobeandmail.com/politics/article-google-to-ban-political-ads-ahead-of-federal-election-citing-new/ Last access: March/11/2020.

[105] S. Shukla, A Better Way To Learn About Ads On Facebook, About Facebook, 03/28/2019, retrieved from https://about.fb.com/news/2019/03/a-better-way-to-learn-about-ads/ Last access: March/11/2020.

[106] J. Canfield, Increasing Transparency Through Advertiser Identity Verification, Google, 04/23/2020, retrieved from https://blog.google/products/ads/advertiser-identity-verification-for-transparency/ Last access: August/ 27/2020.

[107] A. Stamos, An Update On Information Operations On Facebook, About Facebook, 09/06/2017, retrieved from https://about.fb.com/news/2017/09/information-operations-update/ Last access: March/4/2020.

[108] V. Gadde; Y. Roth, Enabling Further Research Of Information Operations On Twitter, Twitter Blog, 10/17/2018, retrieved from https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html Last access: March/11/2020.

[109] A. Woodford, The Hunt For False News, About Facebook, 10/19/2018, retrieved from https://about.fb.com/news/2018/10/inside-feed-hunt-false-news-october-2018/ Last access: March/11/2020.

ment suggests that Facebook's preferred tools for identifying false information are machine learning and reports from external verifiers.[110]

These actions account for a change in the way companies see themselves and show themselves to their different audiences.

Policy changes directly impact the definition and implementation of content policies, community guidelines, or internal rules that affect how companies make moderation decisions.[111] In this case, these are changes to the content that the platforms consider acceptable or not.

In April 2017, for example, Facebook issued a document in which it defined some key concepts related to "information operations" that impacted its internal policies.[112] In June 2019, Twitter updated its rules and simplified them to make them easier to understand.[113] By September 2019, however, Twitter and Facebook announced more radical changes that show their approach to the issue and — to a lesser extent — the scale of the challenge. Thus, Twitter announced the total suspension of political advertising, under the premise that the "reach" of political messages has to be "earned, and not bought."[114]

> "At the end of last year, Twitter decided to ban the promotion of political content worldwide. We made this decision based on our belief that the reach of political messages should be earned, not bought. We define content of a political nature as that which refers to a candidate, political party, elected or appointed government official, election, referendum, measure submitted to a vote, law, regulation, directive, or court ruling. Ads that contain references to political content, including requests for votes, requests for financial support, and advocacy for or against the types of po-

[110] *Ibid.*

[111] But these changes are not necessarily translated, automatically, into formal changes to the rules. *See* on this point, T. Gillespie (Custodians of the Internet, cit.), 66 ("Though Facebook's policy was instituted in December 2016, as of October 2017, no new wording has been added to the community standards document that could reasonably refer to this change. One could argue that Facebook does not prohibit fake news, it only labels it; another possibility is that in some cases policies are instituted, but their addition to community guidelines documents lags"). The rules at the time of writing this paper (July 2020) include "fake news" as banned content. *See* https://www.facebook.com/communitystandards/false_news. On this issue, see the CELE and Linterna Verde, at https://letrachica.digital/

[112] J. Weedon; W. Nuland; A. Stamos, "Information Operations And Facebook," cit.

[113] D. Harvey, Making Our Rules Easier To Understand, Twitter Blog, 06/07/2019, retrieved from https://blog.twitter.com/en_us/topics/company/2019/rules-refresh.html Last access: March/11/2020.

[114] (5) Jack On Twitter: «we've Made The Decision To Stop All Political Advertising On Twitter Globally. We Believe Political Message Reach Should Be Earned, Not Bought. Why? A Few Reasons…» / Twitter, Twitter, 10/30/2019, retrieved from https://twitter.com/jack/status/1189634360472829952 Last access: March/11/2020.

litical content listed above are prohibited under this policy. We also will not allow advertisements of any kind from candidates, political parties, or elected or appointed government officials. Anyone using the platform can report ads that go against this policy. The reporting mechanism through the application itself is available globally and our users can report through this channel. We are working together with electoral authorities around the world to ensure an open dialogue that contributes to the identification of advertisements that violate this policy."[115]

For its part, Facebook announced that it would no longer verify political information, a surprise action that implied a change of direction that the company explained in terms of freedom of expression.[116] In September 2020 it introduced another significant change: following Twitter, Facebook announced that it would not accept political ads in the week leading up to the election.[117]

Three recent events have prompted platforms to once again modify their approaches to the phenomenon of disinformation. This refers to the global pandemic caused by SARAS-CoV-2 and the COVID-19 disease, the 2020 Black Lives Matter movement protests, and the presidential elections in the United States. These events revealed the difficulty of moderating content in a satisfactory way for all the actors who demand improvements in this regard.[118]

---

[115] Hugo TwitterGov, "*Calidad de la información durante elecciones*" [Quality of information during elections], cit., p. 4.

[116] N. Clegg, Facebook, Elections And Political Speech, About Facebook, 09/24/2019, retrieved from https://about.fb.com/news/2019/09/elections-and-political-speech/ Last access: March/11/2020 («We don't believe, however, that it's an appropriate role for us to referee political debates and prevent a politician's speech from reaching its audience and being subject to public debate and scrutiny. That's why Facebook exempts politicians from our third-party fact-checking program. We have had this policy on the books for over a year now, posted publicly on our site under our eligibility guidelines»).

[117] About Facebook, "New Steps To Protect The Us Elections", cit.

[118] On this issue, see A. Wiener (Trump, Twitter, Facebook, And The Future Of Online Speech, cit.); D. Alba; K. Conger; R. Zhong ("Twitter Adds Warnings To Trump And White House Tweets, Fueling Tensions," cit.); M. Murphy («Facebook Shouldn't Be The Arbiter Of Truth,» Zuckerberg Tells Fox News, MarketWatch, 05/28/2020, retrieved from https://www.marketwatch.com/story/facebook-shouldnt-be-the-arbiter-of-truth-zuckerberg-tells-fox-news-2020-05-27 Last access: July/15/2020). In June 2020, Facebook also announced that it would start "labeling" political content, similarly to Twitter. Cf. M. Zuckerberg (Mark Zucerkberg's Announcement, Zuckerberg's Facebook Wall, 06/26/2020, retrieved from https://www.facebook.com/zuck/posts/10112048980882521 Last access: July/15/2020.) ("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. We will soon start labeling some of the content we leave up because it is deemed newsworthy, so people can know when this is the case. We'll allow people to share this content to condemn it, just like we do with other problematic content, because this is an important part of how we discuss what's acceptable in our society – but we'll add a prompt to tell people that the content they're sharing may violate our policies. To clarify one point: there is no newsworthiness exemption to content that incites violence or suppresses voting. Even if a politician or government official says it, if we determine that content may lead to violence or deprive people of their right to vote, we

The pandemic caused by SARS-CoV-2 brought about another source of pressure, in this case driven by global, regional, and national health authorities. And the platforms seem to have reacted to the new scenario by proactively publishing reliable information about the pandemic.[119]

> "For example, Facebook established a 'COVID-19 Information Center', sharing official information and other credible media outlets. This center promoted various pieces of content about the pandemic, including some that were confusing. YouTube began directing people who watched videos related to the virus to 'get the latest Coronavirus data' from official sources such as the World Health Organization and national governments. Twitter, Instagram, and TikTok took similar actions."[120]

This type of practice has not been without significant challenges. On the one hand, not all governments are reliable sources of news or information, even in contexts such as the pandemic. On the other hand, as D'Urso points out, government statements do not necessarily receive the "clicks" necessary to reach the population, especially when they compete with content that appeals to emotional reactions from the readers, such as the content that usually makes up disinformation campaigns.[121] As we have seen, recommendations on the use of drugs that are not proven to be effective, home remedies or— more simply— the underestimation of contagion risks have been common in the Americas in recent months. The disinformation about the pandemic and its supposed special characteristics (like clearer risks and supposed facility to demonstrate the falsehood of claims) may have led the platforms to "do more" about this type of disinformation than for the electoral issues.[122]

However, the pressure to act was not limited to health issues and continued to grow throughout 2020. In July 2020, for example, powerful advertisers an-

---

will take that content down. Similarly, there are no exceptions for politicians in any of the policies I'm announcing here today")

[119] J. D'Urso, *How The Coronavirus Pandemic Is Changing Social Media*, Reuters Institute for the Study of Journalism, 07/06/2020, retrieved from https://reutersinstitute.politics.ox.ac.uk/risj-review/how-coronavirus-pandemic-changing-social-media Last access: July/10/2020.

[120] *Ibid.*

[121] Cf. V. Bakir; A. McStay, «Fake News And The Economy Of Emotions: Problems, Causes, Solutions», Digital Journalism, vol. 6, 2, 2018, retrieved from https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1345645 («…the fakes news problem concerns the *economics of emotion* … emotions are leveraged to generate attention and viewing time, which converts to advertising revenue…»).

[122] J. D'Urso, How The Coronavirus Pandemic Is Changing Social Media, cit. («During the pandemic, social media companies have shown some signs of going further than before when it comes to removing content…»).

nounced a boycott of Facebook until it takes more decisive action on problematic forms of speech, especially "hate speech."[123] The suspension — for the moment, *sine die* — of Donald Trump's accounts from all three platforms marks the most profound and radical change yet.

Undeniably, one of the main problems of content moderation and self-regulation has to do with the malleability of the criteria and the significant fact that these criteria are often not expressed accurately in the terms and conditions or the community guidelines. To address this difficulty, various initiatives have sought to follow up on these complex legal documents. For example, CELE and Linterna Verde have recently launched LetraChica, the first initiative to address this issue in Latin America.[124]

### *4. Transparency and public relations actions*

The evolution of policies, self-regulation rules, and content moderation has been accompanied by increasing demands for transparency. The three companies under analysis publish "transparency reports" that show intense moderation activity, although the level of detail they offer and the formats in which the information is provided make it difficult to analyze the information.[125]

In electoral matters, as an alternative measure to moderation, the companies gradually took steps to increase the levels of transparency regarding advertisements:[126] indicating the origin of the ad, but also offering context information, the general scope of the campaign, amounts invested, and so on.[127] Facebook announced changes and expansions to this policy on several occasions.[128] Almost

[123] «Mark Zuckerberg: Advertisers' Boycott Of Facebook Will End "soon Enough,"» The Guardian, 7/2/2020, retrieved from https://www.theguardian.com/technology/2020/jul/02/mark-zuckerberg-advertisers-boycott-facebook-back-soon-enough Last access: July/15/2020.

[124] Retrieved from https://letrachica.digital/

[125] *See* https://transparency.facebook.com/, https://transparency.twitter.com/, https://transparencyreport.google.com/?hl=es

[126] P. Kafka, *Mark Zuckerberg*, cit.

[127] This action had various consequences, as Facebook extended its policy to all ads, and not just those of a political nature in October 2017 Cf. R. Goldman (Update On Our Advertising Transparency And Authenticity Efforts, About Facebook, 10/27/2017, retrieved from https://about.fb.com/news/2017/10/update-on-our-advertising-transparency-and-authenticity-efforts/ Last access: 6/March/2020.)

[128] R. Goldman, Making Ads And Pages More Transparent, About Facebook, 04/06/2018, retrieved from https://about.fb.com/news/2018/04/transparent-ads-and-pages/ Last access: March/9/2020 (in April 2018); K. Harbath, Updates To Ads About Social Issues, Elections Or Politics In The Us, About Facebook, 08/28/2019, retrieved from https://about.fb.com/news/2019/08/updates-to-ads-about-social-issues-elections-or-politics-in-the-us/ Last access: March/11/2020 (in May, about elections); R. Leathern, Shining A Light On Ads With Political Content, About Facebook, 05/24/2018, retrieved from https://about.fb.com/news/2018/05/ads-with-political-content/ Last access: March/11/2020; K. Walker, Supporting Election

simultaneously, Twitter announced a similar policy, through the creation of the Advertising Transparency Center.[129] For its part, Google adopted a stricter transparency policy for the 2018 mid-term elections[130] and launched a website containing aggregated information on political ads, restricted to the United States, India, the United Kingdom, and the European Union.[131]

Along the same lines, they implemented collaborations with electoral authorities, offering dialogue in real-time and some degree of openness about what was happening on their platforms. Facebook announced the expansion of this policy in September 2017[132] and showed the results of that collaboration concerning the German elections.[133] Similarly, Twitter stated that partnerships with electoral authorities are a constant for the company when it comes to reacting to specific electoral situations.[134] Facebook announced "special measures" for the 2020 elections.[135]

---

Integrity Through Greater Advertising Transparency, Google, 05/04/2018, retrieved from https://blog.google/outreach-initiatives/public-policy/supporting-election-integrity-through-greater-advertising-transparency/ Last access: March/9/2020.

[129] B. Falck, New Transparency For Ads On Twitter, Twitter, 10/27/2017, retrieved from https://blog.twitter.com/en_us/topics/product/2017/New-Transparency-For-Ads-on-Twitter.html Last access: March/6/2020. The Center is available at https://ads.twitter.com/transparency.

[130] K. Walker, Supporting Election Integrity Through Greater Advertising Transparency, cit.

[131] Google, Political Advertising on Google - Google Transparency Report, Google transparency report, 08/15/2018, retrieved from https://transparencyreport.google.com/political-ads/region/US Last access: March/11/2020.

[132] P. Kafka, *Mark Zuckerberg*, cit.

[133] R. Allan, Update On German Elections, Über Facebook, 09/27/2017, retrieved from https://about.fb.com/de/news/2017/09/update-zu-den-wahlen/ Last access: March/4/2020 («Facebook co-operated closely with German authorities, such as the Federal Office for Information Security (BSI). We provided training for members of Parliament and candidates on online security issues. We also set up a dedicated reporting channel for the BSI for issues related to security and authenticity in the context of the federal elections»).

[134] Twitter public policy, *Update*, cit. («We engage with national election commissions regularly and consistently bolster our security and agent review coverage during key moments of election cycles around the world. We will continue to do this, and expand our outreach so that we've got strong, clear escalation processes in place for all potentialities during major elections and global political events»); Hugo TwitterGov, "*Calidad de la información durante elecciones*" [Information quality during elections], cit., P. 3 («*En este sentido, de manera similar a los modelos que hemos puesto en práctica para elecciones recientes alrededor del mundo, como es el caso de Estados Unidos, Brasil y México, hemos implementado un equipo interno e interdisciplinario para liderar nuestros esfuerzos en materia de integridad electoral. A través de nuestras propias herramientas internas, este equipo trabaja proactivamente para evaluar tendencias, atender casos que son escalados por nuestros aliados en la materia e identificar posibles amenazas provenientes de actores maliciosos. Respecto de colaboraciones específicas con autoridades electorales y actores similares en Latinoamérica, les invitamos a leer nuestro blog referente a las Elecciones en México a manera de ejemplo. Esfuerzos similares fueron de igual forma implementados en Colombia y Argentina. Asimismo, dada la similitud entre esta pregunta y la pregunta número seis, en la respuesta a esta última ahondaremos más al respecto*» [In this sense, similar to the models we have put into practice for recent elections around the world, such as the United States, Brazil and Mexico, we have implemented an internal and interdisciplinary team to lead our efforts in the field of electoral integrity. Through our own internal tools, this team works proactively to assess trends, address cases that are escalated by our allies in the matter, and identify possible threats from malicious actors. Regarding specific collaborations with electoral authorities and similar actors In Latin America, we invite you to read our blog regarding the Elections in Mexico as an example. Similar efforts were implemented in the same way in Colombia and Argentina. Also, given the similarity between this question and question number six, in the answer to the latter we will delve more on it]).

[135] About Facebook, "New Steps To Protect The Us Elections," cit.; G. Rosen, Helping To Protect The 2020 Us Elections, About Facebook, 10/21/2019, retrieved from https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/ Last access: March/11/2020.

Furthermore, during this period companies were summoned to publicly explain their businesses, models, policies, and standards in different forums and were explicitly summoned to formal meetings with public officials (e.g. United States, United Kingdom). Citations from the Congress of the United States seem to be especially significant. In September 2017, The Hill reported that Facebook, Google, and Twitter would be called to testify before the Senate Intelligence Committee on the issue of Russian interference in the 2016 election.[136] In this context, Twitter announced meetings and pointed out obstacles to dealing with fake accounts and bots.[137] In November 2017, the companies announced that they would collaborate with the United Kingdom to investigate possible Russian interference in the referendum for that country's remaining in the European Union in 2016.[138] In September 2018, Facebook's Sheryl Sandberg and Twitter's Jack Dorsey testified again before the US Senate Intelligence Committee.[139] It is interesting to point out that there is an implied questioning of targeted advertisements that pushes companies to defend them;[140] these positions are seen — for example — in the platforms' "principles" regarding their advertising practices.[141]

Finally, at the same time as the development of state monitoring bodies, the companies have developed actions and created procedures and mechanisms with different degrees of openness to third parties (specialists, NGOs, academics, etc.) to generate policies and implementation criteria that satisfy users and regulators. In April 2018, Facebook published its internal moderation criteria and began holding stakeholder meetings on this subject; and expanded its internal appeals

[136] J. Byrnes (Senate Panel Invites Facebook, Google To Testify In Russia Probe, The Hill, 09/27/2017, retrieved from https://thehill.com/homenews/senate/352743-senate-panel-invites-facebook-to-testify Last access: May/7/ 2020.). At the time, Facebook announced that it would cooperate with the authorities. *See* C. Stretch (Facebook To Provide Congress With Ads Linked To Internet Research Agency, About Facebook, 09/21/2017, retrieved from https://about.fb.com/news/2017/09/providing-congress-with-ads-linked-to-internet-research-agency/ Last access: 6/March/2020.)

[137] Wired, «Twitter Will Meet With Senate Intelligence Committee On Russia», Wired, 09/28/2017, retrieved from https://www.wired.com/story/twitter-senate-committee-russia-bots/ Last access: May/7/ 2020.

[138] M. D. Stefano, Facebook And Twitter Say They'll Cooperate With The Uk Inquiry Into Russian Meddling In Brexit, BuzzFeed, 11/29/2017, retrieved from https://www.buzzfeed.com/markdistefano/facebook-and-twitter-say-theyre-ready-to-co-operate-with Last access: March/6/2020.

[139] US Senate Intelligence Committee, Hearings | Intelligence Committee, US Senate, 09/05/2018, retrieved from https://www.intelligence.senate.gov/hearings/open-hearing-foreign-influence-operations%E2%80%99-use-social-media-platforms-company-witnesses Last access: May/7/ 2020.

[140] E. Schrage, "Hard Questions," cit. («These are worthwhile uses of ad targeting because they enable people to connect with the things they care about. But we know ad targeting can be abused, and we aim to prevent abusive ads from running on our platform. To begin, ads containing certain types of targeting will now require additional human review and approval»).

[141] *See*, e.g., Facebook's Principles in R. Goldman, Our Advertising Principles, About Facebook, 11/27/2017, retrieved from https://about.fb.com/news/2017/11/our-advertising-principles/ Last access: March/6/2020.

procedure on its decisions.[142] By then, Twitter had already launched its Trust and Safety Council and recently it announced an expansion of its powers in line with Facebook's announcement.[143] By late 2018, Facebook also announced the creation of a council of specialists that would help the company make good decisions regarding moderation, compatible with the principles of freedom of expression that Zuckerberg publicly emphasized at various times. The Oversight Board was finally presented publicly in May 2020.[144]

This type of action seems to respond to the growing demands for "accountability" in terms of content moderation and is a step in the direction of turning large platforms into "public forums."

## *C. Analysis: good practices, problems, and challenges*

The phenomenon of disinformation arose as a problem after the 2016 presidential election in the United States. The big Internet platforms have had to deal with it ever since. In its previous report, CELE detected that as of December 2017 the platforms had taken actions that were "in the preliminary phase."[145]

[142] M. Bickert (Publishing Our Internal Enforcement Guidelines And Expanding Our Appeals Process, About Facebook, 04/24/2018, retrieved from https://about.fb.com/news/2018/04/comprehensive-community-standards/ Last access: March/9/2020.). *Community standards* can be found in https://www.facebook.com/communitystandards/. In May 2018, Facebook published a report on cases to show how its criteria operate in practice. *See* G. Rosen (Facebook Publishes Enforcement Numbers For The First Time, About Facebook, 05/15/2018, retrieved from https://about.fb.com/news/2018/05/enforcement-numbers/ Last access: 9/March/2020.)

[143] N. Pickles, Strengthening Our Trust And Safety Council, Twitter Blog, 12/13/2019, retrieved from https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-trust-and-safety-council.html Last access: August/ 27/2020.

[144] (Establishing Structure And Governance For An Independent Oversight Board, About Facebook, 09/17/2019, retrieved from https://about.fb.com/news/2019/09/oversight-board-structure/ Last access: March/11/2020.); *see* https://www.oversightboard.com/.

[145] C. Cortés; L. Isaza, "*Noticias falsas en Internet: La estrategia para combatir la desinformación*" [Fake news on the Internet: The strategy to combat misinformation], cit., P. 26.

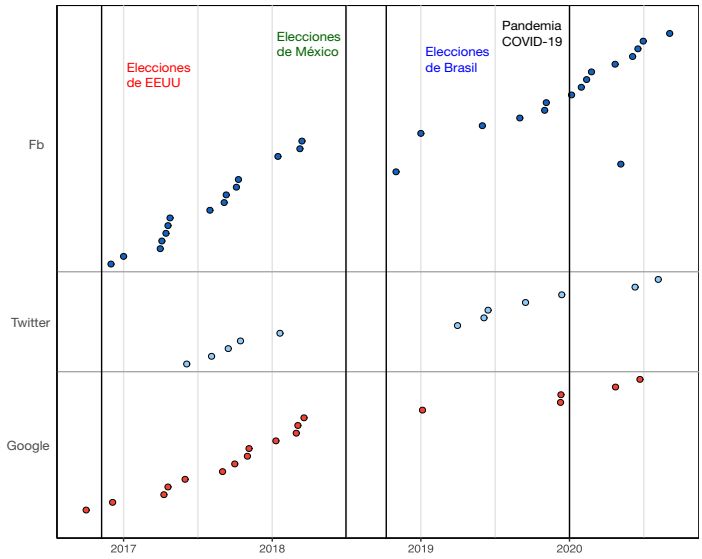


**Figure 3** - Actions Timeline

We should go over them to get a complete picture of the findings that this report seeks to update.

- *Facebook*. It acted based on a definition of "information operations" as the actions carried out by organized actors to distort national or foreign political sentiment, and they are divided into fake news, disinformation, and fake accounts. The 2017 report identified two types of actions: positive ones focused on raising user awareness, and negative ones that were "technical" solutions to the challenge of misinformation. The latter included (a) reporting, fact-checking, and warnings; (b) changes in the newsfeed and measures to counteract false amplification.
- *Google*. In addition to awareness-raising actions, the report disclosed changes in search services, especially linked to highlighting "quality" content and verification services. The report noted, however, that "Google is limited to highlighting third-party verifications based on their criteria, even if the conclusions are different."[146] The point seems significant because a review of the current search system from Argentina did not confirm that any classification system that highlights the verification of contents has a special ranking in the usual search results. On the other hand, the report also identified the possibility of reporting predictions, an option that is available in Latin America but not linked to misinformation.

[146] *Ibid.*, p. 17.

- *Twitter*. Actions identified in 2017 included reducing the visibility of tweets and potential spam accounts while under investigation, suspending accounts, and detecting applications that abuse their API.

At the same time, the report outlined challenges and possible solutions.[147] We should mention them: the report pointed out problems of scale and time and detailed that disinformation could not be dealt with only with automated mechanisms; therefore it would require an unfeasible amount of human resources. Likewise, the report highlighted that making "good" moderation decisions on this front required the intervention of people with the necessary judgment capacity to rule, with some degree of certainty, that we are in fact facing a problematic disinformation campaign. Furthermore, the report noted that such structures are difficult to imagine in secondary markets, like Latin America. The report also denounced the possibility that some of the proposed solutions were subject to the dynamics of "unexpected consequences:" that — for example — the verification of information would have the opposite of the desired effect, such as reinforcing erroneous beliefs or conspiracy theories which often bolster these beliefs.

Finally, we should underline an additional point: the 2017 report stated that "the drafting of this document had a constant difficulty: understanding if the large number of decisions and actions covered in the media and announced by companies were effectively implemented and to what degree."[148] This difficulty persisted during the drafting of this follow-up document: of the 61 most relevant actions identified and outlined in this document, the effective implementation of 28 of them could not be verified and most of the actions verified correspond to the more visible actions, such as those to support journalism, partnerships with verifiers, etc. (Figure 4).

[147] *Ibid.*, p. 21-24.

[148] *Ibid.*, p. 23.

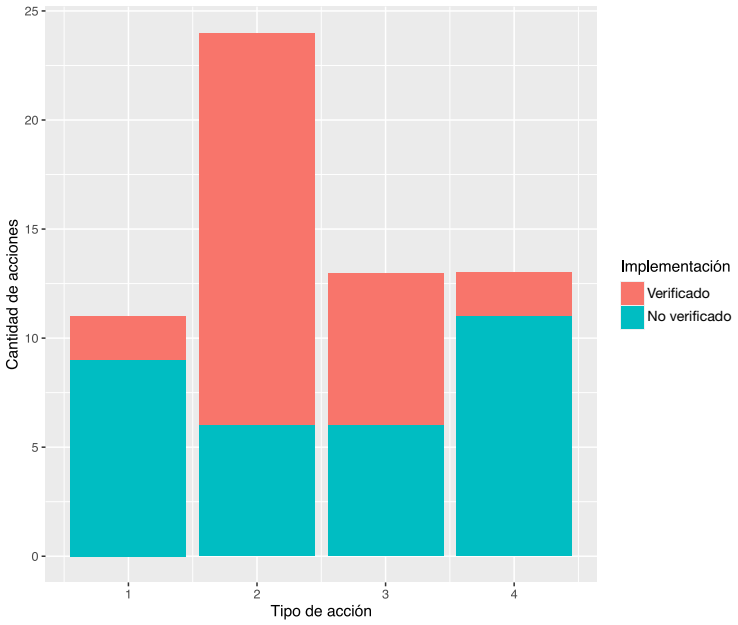


**Figura 4** - Types of actions and implementation status

An analysis of what has happened in recent years suggests that the patterns identified in the previous report are still valid, but there have appeared some new trends. Thus, for example, the creation of internal bodies in charge of influencing moderation policies appears as one of the most significant innovations. Facebook's Oversight Board and-to a lesser extent-the Trust and Safety Council of Twitter are new attempts by platforms to increase accountability for their actions and content moderation policies, although both actions seem to follow the model of the "public forum" that slowly seems to prevail regarding this type of company.

On the other hand, recently certain trends seem to have become stronger: in addition to safety awareness actions and the promotion of more information, the platforms disclosed here have opted for practices of "contextualizing" the information they present to users, increasingly thanks to content recommendation algorithms. The practice of labeling seems to have grown: in addition to labeling in relation to content verification, both Twitter and Facebook have labeled posts from important public actors as containing false information, government media from foreign countries, and so on. Companies seem to continue to invest in artificial intelligence solutions for their moderation policies: a sufficiently well-trained algorithm could identify false information and act accordingly, either by labeling it as such, establishing some kind of prior warning, reducing its circulation in the network, and so on. This reliance on algorithms, however, could be

excessive. As the companies themselves recognize, the line that separates disinformation from satire or opinion protected by freedom of expression is very thin and - for the moment - it does not seem possible that a purely technical solution could be effective on its own, despite the fact that as a consequence of the pandemic, platforms have fast-tracked the adoption of this type of tool.[149]

Finally, some general trends appear to be gaining strength. The companies analyzed in this study — especially Facebook and Twitter — appear to have embraced a firm exercise of their moderation powers. The latest actions identified seem to support the hypothesis of context: identifying the location of information, labeling government media, blocking advertisements from foreign media, and suspending political advertising in the case of Facebook, or labeling political content, suspending political advertising, or labeling government media in the case of Twitter show actions from the last year in that direction. They occur in a context of growing pressure on the business model posed by social media, various regulatory threats, and even investigations by regulatory authorities for anti-competitive practices.[150]

The string of actions in recent years shows companies that are very active in the face of a phenomenon that causes concern and about which little is known. This makes it difficult to reach definitive conclusions, but some relevant questions can be raised at least tentatively. For example, what is the impact of these actions in Latin America? The usual response from platforms — consulted for this document, but also expressed in the framework of other "multi-stakeholder" initiatives —is that company policies are global. However, there were many actions identified whose implementation in Latin America could not be verified, as shown in Figure 3. Clearly, there is a gap between global announcements and their implementation in Latin America.[151]

On the other hand, are there good practices that result from the actions identified? The answer to this question can also only be tentative. If by good practic-

[149] C. Newton, The Coronavirus Is Forcing Tech Giants To Make A Risky Bet On Ai, The Verge, 03/18/2020, retrieved from https://www.theverge.com/interface/2020/3/18/21183549/coronavirus-content-moderators-facebook-google-twitter Last access: August/ 27/2020.

[150] C. Newton, The Antitrust Case Against Google Is Gathering Steam, The Verge, 07/15/2020, retrieved from https://www.theverge.com/interface/2020/7/15/21324105/google-antitrust-california-search-youtube Last access: September/22/2020; A. Satariano, «Facebook Loses Antitrust Decision In Germany Over Data Collection», The New York Times, 06/23/2020, retrieved from https://www.nytimes.com/2020/06/23/technology/facebook-antitrust-germany.html Last access: September/22/2020.

[151] *See*, in this regard, Twitter's response to the inquiry regarding the possibility of reporting content as false for electoral reasons. Cf. Hugo TwitterGov ("*Calidad de la información durante elecciones*" [Information quality during elections], cit.), 4-5.

es we mean actions that are respectful of current human rights standards, the answer could be positive: evidently, there are more or less problematic actions within that parameter. But if what we are looking for are effective actions, the answer is more elusive because in that light these actions are difficult to evaluate: there are no impact studies, there is not enough access to the necessary information by companies and there does not seem to be any concern for gauging results. Rather, viewed as a whole, the actions seem to follow one another in an effort to respond to increasing pressure and unrest.

A hypothesis that explains the number of actions undertaken is related to the following fact: within the multiple points of pressure that companies bear, the threat of regulatory action is always present. By showing proactivity and receptivity to the demands of powerful actors, the companies under review outweigh the arguments against these efforts. On the other hand, this permeability is valued by state agents who find — through informal influence — an effective channel to exercise some control over the circulation of information on the Internet. Concentration facilitates this result: large platforms are *de facto* necessary intermediaries that become influential control nodes.

This scenario raises big issues regarding the role of intermediaries in the circulation of information on the Internet. If this question had a definite answer in 1996 through the guarantees of section 230 of the Communications Decency Act, it is clear that this is no longer the case. The role and liability of large Internet platforms are in full dispute throughout the world and some type of regulatory innovation will be imposed in the not too distant future. However, we still do not know the form that regulation will take: whether we will continue with self-regulatory actions like the ones we have seen so far — whose intensity and variety seem to have increased during the pandemic — or if a traditional regulation will prevail along with states with the capacity to influence what platforms do, such as the United States or the European Union. The issue will probably be addressed with a variety of approaches, as suggested by Marsden, Meyer, and Brown: controlled self-regulation, a formalized self-regulation, or co-regulation seem the most plausible scenarios.[152] In this sense, what we are currently seeing is — possibly — a slow shift in the public debate in that direction (Figure 4).

[152] C. Marsden; T. Meyer; I. Brown, «Platform Values And Democratic Elections: How Can The Law Regulate Digital Disinformation?», Computer Law & Security Review, vol. 36, 2020, retrieved from http://www.sciencedirect.com/science/article/pii/S026736491930384X.

Some signs support this hypothesis. Indeed, although the legal basis for moderation actions continues to be the property rights of companies and the power derived from defining what content is acceptable or not, a large part of these decisions is based on legal requirements of relevant countries, such as with the unproblematic cases of child pornography or in the most questionable cases of "hate" speech, thanks to the regulations in force especially in European countries. In the Marsden *et al*. option menu we are currently between the second stage of self-regulation and the third, where external actors begin to influence, for example, the GNI audits or the European Commission Code of Practice on Disinformation.[153]

This possible future scenario will involve a questioning of the principles of freedom of expression that, until now, have defined this issue. The non-liability of intermediaries for the content produced by third parties operated as a viable principle within the framework of a decentralized network. The processes of concentration and centralization that characterized the pessimistic turn regarding the Internet cast doubt on whether this principle is sufficient to address multiple current problems, from disinformation to the algorithmic creation of new content. In any case, this discussion will take place in a complex scenario, possibly in regulatory models different from the "pure" models that we have imagined so far.[154]

In this scenario, Latin American countries will play a secondary role, and this reveals the persistence of a structural challenge in everything related to the issue of regulation of large Internet platforms from peripheral countries: not all countries have the same power to influence how these companies act.[155] In this sense, it is difficult to imagine that the countries of the region can radically modify the major trends that can be observed today, but perhaps they can have an impact. In recent years we have seen that the platforms have cited some local experiences as virtuous dynamics, such as the collaborations between the platforms and the electoral authorities in recent electoral processes in Mexico and Brazil. So far, we have not seen "traditional" regulatory initiatives introducing noteworthy innovations; on the contrary, the last experience in this regard was the Brazilian bill

[153] European Commission, «EU Code Of Practice On Disinformation». European Commission, Brussels. September 2018; C. Marsden; T. Meyer; I. Brown, "Platform values and democratic elections," cit., P. 12.

[154] On this issue, *see* the Abbott and Snidel model; cf. K. Abbott; D. SnidalWalter Mattli, Ngaire Woods (eds.) («The Governance Triangle: Regulatory Standards Institutions And The Shadow Of The State», in The politics of global regulation, Princeton University Press, Princeton, 2009.)

[155] This is a global problem, linked to the role that transnational companies play in respecting or violating human rights in peripheral countries. On this subject, *see* J. Ruggie ("Protect, Respect And Remedy: A Framework For Business And Human Rights." Human Rights Council. Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, Geneve. A/HRC/8/5. April 7, 2008.), Par. 14

on *Liberdade, Responsabilidade e Transparência Digital na Internet*, which was unanimously rejected by civil society in the region for affecting the fundamental principles of freedom of expression that, for the moment, control the type of acceptable responses to these problems.

From this perspective — that is, from the point of view of the current standards of freedom of expression — the various actions of the platforms easily fit within an imaginary spectrum of what is acceptable. The actions included in category (1) are the least problematic and respect the fundamental principle that the best solution to the abuses of freedom of expression is *more freedom of expression*.[156] Errors are answered with corrections; false information is fought with true information, and so on. Moderation actions that are limited to providing more information, in principle, are also part of the same normative conclusion. The outlook begins to change, however, in cases in which these moderation actions impact, for example, the reach of content that may have been flagged as problematic and whose recommendations were reduced.[157] These types of effects, generally difficult to identify, impact how information circulates and turns intermediaries into true "curators" of public debate, which — as we have seen — is the role that they are assuming increasingly. In these cases, those actions that have an impact on the public debate become more problematic and must be submitted to the normative analyzes required by the Inter-American system, which are fundamentally based on using transparent criteria, consistently, and guaranteeing due process for those users whose contents are affected.[158]

Transparency in moderation actions should improve substantially. Transparency reports are difficult to analyze and provide information in hard-to-read or overly aggregated formats. As a general rule, we are unaware of specific cases of moderation: to bridge this gap, platforms should be open to studies by independent parties and academics.[159] On the other hand, moderation rules should also be clearer and their application should be consistent: the platforms analyzed

[156] IACHR, «*Marco Jurídico Interamericano del Derecho a la Libertad de Expresión*» [The Inter-American Legal Framework regarding the Right to Freedom of Expression] Office of the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights. OEA/Ser.L/V/II CIDH/RELE/INF. 2/09. December 30, 2009. Par. 108, *et seq*.

[157] YouTube, "Continuing Our Work To Improve Recommendations On Youtube", cit.

[158] IACHR, «*Libertad de Expresión e Internet*» [Freedom of Expression and the Internet]. Inter-American Commission on Human Rights, Washington D.C. 2013. Par. 55; «*Estándares Para Una Internet Libre, Abierta e Incluyente*» [Standards for a Free, Open and Inclusive Internet]. Office of the Special Rapporteur for Freedom of Expression of the IACHR, Washington D.C. INF.17/17. 2017. Par. 87.

[159] There is an initiative that leans in this direction: Social Science One. *See* https://socialscience.one/

in this document are not thoroughly evaluated by independent actors.[160] These needs, which we barely point out here, are related to the curation role that platforms have adopted in the complex and changing regulatory scenario that we explained in previous paragraphs.

We still do not know if that role will be accepted by the regulation that will eventually be introduced. We imagine other possible worlds, in which that curating role is rejected as unnecessary: in a less concentrated and more decentralized network, similar to the original model, that role of curation would be unfeasible and inefficient. Information would circulate in a somewhat more chaotic way and there would be no simple way to exercise the control that is demanded of these central actors today.

However, that future seems distant. Although there are investigations from authorities in defense of competition around the world (especially in Europe), we see incentives aligned to maintain and safeguard the central role that the actors analyzed have achieved. It is easier for states to control the circulation of information when there are central actors with control capacity than when these actors are absent, there are too many, or do not concentrate significant portions of the traffic. The coincidence of these interests with the private interests of some of the most powerful corporations in the world suggests that the current regulatory path aims at accepting these platforms (at this level of concentration). As long as this characteristic of the Internet persists, and to the extent that various actors — international organizations, NGOs, research centers, and states — continue to demand concrete actions from the platforms against threats to democracy perceived as serious, greater transparency regarding the criteria used for moderation seems like a reasonable demand. In this context, platforms assume a role that is increasingly similar to that of "public forums" that are subject to criteria and standards not entirely under their control. That end-point is not inescapable, and it may not be desirable. But it seems that we are heading in that direction.

[160] On this issue, *see* tos;dr, in https://tosdr.org/