Fake news on the Internet: actions and reactions from three platforms

Submission to the Special Rapporteur for Freedom of Opinion and Expression, Irene Kahn, Feb 2021

CELE

The Center for Studies on Freedom on Freedom of Expression and Access to Information is an academic research center affiliated with Universidad de Palermo in Argentina. The Center provides technical, legal analysis on issues affecting this fundamental right, and since 2012 has been studying freedom of expression on the Internet as a specific research area. The Center is a leading voice on the promotion and protection of freedom of expression nationally, regionally and internationally.

This submission was prepared in response to the public call for input published by the Special Rapporteur. Per the call, "The Special Rapporteur will seek to clarify how human rights law applies to disinformation, identify key issues that would benefit from further consideration by the Human Rights Council and formulate recommendations to States and other stakeholders on the best way to tackle disinformation whilst protecting the right to freedom of opinion and expression."

CELE has studied and published on the issue of disinformation from various points of view. The Center runs a repository of laws and bills of law encompassing 9 countries in Latin America that tracks, among others, legislative efforts to the deal with the phenomenon; and in 2020 launched Letrachica.digital, a project that tracks changes to terms of service and community guidelines in real time. Both projects seek to understand how public and private regulations and restrictions of content work and impact the exercise of freedom of expression, particularly in Latin America. This submission is based on a research project that traked and documented the different actions taken by companies to address disinformation, especially related to electoral issues. The submission seeks to address point 3 of the questionnaire, which reads:

a. *What policies, procedures or other measure have digital tech companies introduced to address the problem of disinformation?*

Mario Bravo 1050, 7º piso (C1175ABT), CABA, Argentina | TE: (54 11) 5199-4500 int 1213 | Fax: (54 11) 4963-1560 | cele@palermo.edu | **www.palermo.edu/cele**

b. *To what extent do you find these measures to be fair, transparent and effective in protecting human rights, particularly freedom of opinion and expression?*

c. *What procedures exist to address grievances and provide remedies for users, monitor the action of the companies, and how effective are they?*

Further information, research and analysis on this an other issues are available at www.palermo.edu/cele or through email to cele@palermo.edu. We thank the Office of the Special Rapporteur for considering this submission.

Introduction

The manipulation of public opinion through lies has been a constant throughout history. However, disinformation as a modern phenomenon emerged recently in the context of electoral processes that took place over a contaminated public debate, in part by the dissemination of false information on the Internet. Since the election of Donald Trump to the presidency of the United States in 2016, the phenomenon has been recurrent in several electoral processes with varying intensity and impact.

For many observers, online disinformation constitutes a serious threat to the future of democratic systems. Faced with this challenge, states around the globe have reacted in different ways. Some have sought to regulate the practice and impose on intermediary platforms, where disinformation thrives, heavy duties of moderating the content they allow. Others have sought to increase the awareness of the population regarding the existence of disinformation campaigns meant to deceive. Companies, on the other hand, have sought to take action under increasing, and often contradictory, pressures from state institutions, regulators, legislators, NGOs and academics around the world.

This is an executive summary of a report prepared by CELE that sought to track the actions taken by Google, Twitter and Facebook on disinformation, covering the period between 2016 and 2020 with a special view towards Latin America. From our perspective, in order to understand disinformation and the actions taken by these leading intermediary companies it is necessary to understand the temporal dimension within which the recent phenomenon of disinformation arisen. It is also useful to understand how the problem and the answers offered to address it are part of a broader shift in public attitudes towards democracy and the Internet: if in the 1990s the network was seen as a democratizing force that would topple dictatorships through the free flow of information, since the early 2000s a more pessimist approach has been slowly but steadily capturing the heart and minds of

Mario Bravo 1050, 7º piso (C1175ABT), CABA, Argentina | TE: (54 11) 5199-4500 int 1213 | Fax: (54 11) 4963-1560 | cele@palermo.edu | **www.palermo.edu/cele**

regulators around the world. The Internet is no longer perceived as a democratizing force for good—it is often seen as a threat to democracy itself. This change questions the legal standards of intermediary (non) liability that until now have offered a solution to the problem of accountability in decentralized networks. This legal framework seems increasingly untenable.

A. Background and Challenges

In the beginning, being a decentralized network, the Internet was presented as the ideal space for the free flow of information and ideas of all kinds. This initial promise generated some optimism regarding its influence on the future of democracy: the Internet could be a democratizing tool in closed societies, since its decentralized nature would make it difficult for totalitarian governments to control their citizens. 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 is the prime example of such approach. The goal of that regulation was to encourage innovation and facilitate, rather than hinder, the flow of information.

Over time, however, some caution emerged. Authoritarian governments soon learned to use the Internet to increase their control over their citizens. Several authors, starting in the late 1990s warned that the Internet was in a troublesome path that needed to be corrected. This started a slow but steady pessimist turn that would put Internet companies in the wrong side of the good and evil divide, something that went along a pattern of concentration of property. Various new intermediary emerged in those years and became increasingly powerful. Google (1998), Facebook (2004), YouTube (2005) and Twitter (2006) began to concentrate a large part of the traffic and started a process of re-centralization of the network. Their terms of service and community guidelines operate as *de facto* regulations of the information that is reputed acceptable or not on each of these platforms. The power that Internet platforms actually enjoy became, over time, the reason why various actors started to demand greater responsibility and accountability. This has pushed platforms to do much more than simply “facilitate” the free flow of information.

Since the 2016 US presidential election, the main platforms appear to be on the defensive. On a regular and somewhat chaotic basis, they have announced and implemented changes to their platforms aimed at combating the phenomenon of disinformation; they have testified at formational hearings before legislative bodies around the world; they have published studies and provided information as part of their efforts to offer some

Mario Bravo 1050, 7º piso (C1175ABT), CABA, Argentina | TE: (54 11) 5199-4500 int 1213 | Fax: (54 11) 4963-1560 | cele@palermo.edu | **www.palermo.edu/cele**

transparency; they have supported and led programs to strengthen quality journalism and the "verification of information".

Companies are pushed to adjudicate what is true and what is not, a duty that used to fall on citizens in democratic societies. Suddenly, we have incorporated an intermediary actor, central to the flow of information, as an arbiter of truth that cannot be influenced through democratic procedures. This changes pose new challenges: do we maintain our commitment to broad freedom of expression standards and ratify that the best answer to false information is the free competition of the market of ideas? Or should this paradigm be revised? In the second case, what do we replace it with?

B. Actions and Main Findings

The actions we have identified are linked to the temporality of the phenomenon of misinformation and to changes in attitude towards the Internet on the optimism / pessimism axis. They sought to face the demands of users, civil society and governments in relation to the central role that companies acquired in the flow of information on the Internet. In a way, these actions are consequences of their own success. The report found hundreds of actions: some announced and implemented others only partially implemented, and finally an important number of actions that could not be verified at least in Latin America. We classified them in four different categories.

1. *Awareness-raising Actions*. These are aimed at raising awareness about disinformation campaigns or promoting quality journalism as a presumably effective response to the phenomenon. They involve alliances with other actors (e.g., fact-checking organizations), education campaigns, digital and media literacy, etc. Most of these actions took place in the United States, Canada and the United Kingdom, but they have also spread –though with significant limits- to other areas of the world, including Latin America. Alliances with fact-checkers seem to have been a preferred response by platforms to confront the phenomenon of disinformation. This type of alliance does not represent significant problems from the point of view of freedom of expression. However, many challenges remain, such as limited impacts on the less informed or less educated, who—in turn—are more likely to consume false information.

2. *Changes to Code*. These actions modify the code of platforms, changing the recommendation mechanisms and the visibility of content. Platforms are increasingly introducing algorithm-based timelines that—supposedly—seek to bring users more "relevant" content for them, determined by consumption patterns that the platforms themselves record and exploit. Within this category, we found actions aimed at

Mario Bravo 1050, 7º piso (C1175ABT), CABA, Argentina | TE: (54 11) 5199-4500 int 1213 | Fax: (54 11) 4963-1560 | cele@palermo.edu | www.palermo.edu/cele

detecting false information, either through the assistance of users (through e.g. expanding reporting mechanisms) and/or through artificial intelligence. We also found actions aimed at providing more context to information, especially on issues where disinformation was detected and deemed as a problem, such as elections and the COVID-19 crisis that began in late 2019. These actions have faced important challenges due to their context intensive dependency and the issues associated with scale and replicability. Additionally, there are inequalities as to the information available to counter misinformation or conflicting information in different jurisdictions. This in turn creates new challenges based on predictability, and uniformity in the application of social media rules and self-regulation standards.

Companies have also sought to give a more powerful voice to "professional" journalists, media outlets and fact-checkers, prioritizing their content and its visibility. This approach, which turns on platforms own free speech rights, is based on a problematic founding principle though: traditional media is trustworthy and are means to combat disinformation rather than amplifiers of such a phenomenon. As described by multiple researches recently, this is not necessarily true.

Finally, companies have also sought to attack the problem indirectly through restricting "behaviors" (rather than content) that disinformation agents use to advance their campaigns, such as the use of bots and the massive distribution of content in encrypted platforms. This provides scalability and opportunities for automation that other actions don't.

3. *Changes to policy and moderation actions*. This category encompasses the actions aimed at changing the rules that define which speech is acceptable within each platform. These are actions that may have an impact on their business model, for they often involve self-imposed limitations without necessarily changing the code of the platforms. These actions, so far, were informed by companies´ refusal to control the information that their users share through their services, a position based on the optimistic paradigm of the 1990s. However, increasing pressure on companies to exercise their moderation power more decisively has led them in that direction, especially on electoral issues and—after the COVID-19 crisis—on matters of public health. These changes are significant, and show flexible companies, with the ability to adapt their policies to the growing pressures to play a more relevant role in controlling the information that flows through their services. Still, this adaptability and susceptibility to pressure also show the arbitrariness and instability of the rules, the lack of transparency and predictability in its application and the increasing

Mario Bravo 1050, 7º piso (C1175ABT), CABA, Argentina | TE: (54 11) 5199-4500 int 1213 | Fax: (54 11) 4963-1560 | cele@palermo.edu | **www.palermo.edu/cele**

susceptibility of self-regulation to different social and political pressures to the detriment of freedom of expression.

4. *Transparency and public relations*. These are actions aimed at establishing the companies´ position internally and in face of external pressures: on the one hand, the political sectors that can regulate them through legislative changes; on the other, social sectors that exert pressure on them. Advertisers who view certain practices with concern are another group that pays attention to how platform policies evolve. These actions generally show companies willing to address external concerns, and accommodating often contradictory demands. They also show companies taking a stance regarding their role in the free flow of information, often –until now- unwilling to become "arbiters of truth". Furthermore, and under pressure, companies often produced information at the request of authorities, such as e.g. on the influence of Russian intelligence operations on their platforms. Finally, within this category, we should include the creation of procedures and mechanisms to generate adequate implementation policies and criteria that satisfy users and regulators is one of the main novelties in recent years. In April 2018, Facebook published its internal moderation criteria, expanded its internal appeals procedure on its decisions, and promised responses within 24 hours. By then, Twitter had already launched its Trust and Safety Council and recently announced an expansion of its powers in line with the body announced by Facebook. In 2018 Facebook announced the creation of the Oversight Board, a council of notables that would help the company make good decisions regarding moderation, compatible with the principles of freedom of expression, which presented its first decisions in January 2021.

C. A Changing World

The pessimistic turn on the Internet and the *re-centralization* of the network puts into question the legal solution reached so far to deal with content produced by third parties. If non-liability of intermediaries was until a few years ago a matter of course, that no longer seems to be the case. The increasing pressure on big platforms to exercise more vigorously their moderation powers has led them in the path of *public forums*—they no longer seem to enjoy an absolute prerogative to decide which speech is allowed on their services and which is forbidden. Outside criteria, including international human rights standards, seem to increasingly be playing a function.

This happens, however, in a context in which regulatory action by state agents appears only in the form of a threat. Aside from the laws adopted in Germany and France, no country in

Mario Bravo 1050, 7º piso (C1175ABT), CABA, Argentina | TE: (54 11) 5199-4500 int 1213 | Fax: (54 11) 4963-1560 | cele@palermo.edu | www.palermo.edu/cele

the West has addressed the challenge of disinformation through legislative law-making. Our research suggests that companies act under pressure but in the exercise of their own self-regulatory powers. In that sense, our report suggests that the pattern of findings captured in CELE's 2017 report is deepening.

1. Platforms have embraced a more robust moderation role, a development that the COVID-19 crisis seems to have encouraged. In the last two years, companies have moved forward with information localization actions, government media tagging, blocking of foreign media advertisements, suspension of political advertising, the labeling of political content, among others.

   These actions happen within a context of uncertainty. There is still a lot we do not know about disinformation campaigns: how they operate, who are behind them, what are their effects. Platforms' actions are developed in that context. The bet towards the automatic detection of disinformation is bold precisely because it is a difficult issue to identify, that is crossed by deep disagreement in terms of what it is and what the proper reaction of a democratic society should be. The actions that provide more "context" move within the old paradigm of freedom of expression, and for now seem to be more salient than direct acts of censorship, which appear to be measures of last resort such as e.g. the *deplatforming* of Donald Trump in January 2021. Viewed as a whole, the actions our report analyze seem to follow one another somewhat chaotically in an effort to respond to growing pressures and discomforts.

2. The development of corporate bodies more or less "independent" from the companies, like Twitter's *Trust and Safety Council* or Facebook's *Oversight Board* appear as governance innovations that are promising insofar as they are capable of achieving some degree of legitimacy in front of the external actors who exert pressure on companies. It is, however, too soon to tell whether their efforts will be successful.

3. In our analysis several actions announced in the United States or in Europe have not yet been implemented in Latin America. There is an "implementation gap" that is problematic in and of itself.

4. Transparency on moderation as well as other actions is still lacking. 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 except the ones that are covered by mass media. To resolve this gap, platforms should be open to studies by independent actors and academics. On the other hand, the moderation rules should also be clearer and their application should be consistent.

Mario Bravo 1050, 7º piso (C1175ABT), CABA, Argentina | TE: (54 11) 5199-4500 int 1213 | Fax: (54 11) 4963-1560 | cele@palermo.edu | **www.palermo.edu/cele**

This scenario poses profound dilemmas regarding the role of intermediaries in the flow of information on the Internet. The processes of concentration and re-centralization that characterized the pessimistic turn regarding the Internet cast doubt on whether the principle of non-liability of intermediaries is enough to address multiple current problems, from disinformation to the algorithmic creation of new content. Regulatory innovation will occur, but we are less certain about the shape it will take. From our perspective, the current trend of *self-regulation* is not to last long, but traditional, nation-state regulation is not going to be the answer either. With initiatives such as the Global Network Initiative in place, we see a scenario of *co-regulation* such as the one described recently by Marsden, Meyer and Brown, as a much more likely turn of events in the not so distant future.

The *re-centralization* trend may also come under increased scrutiny if current antitrust investigations, both in the United States and in Europe, move forward. In a less-concentrated Internet and a more decentralized network, more similar to the original model, moderation would be less effective. Information would circulate in a somewhat more chaotic way and there would be no simple ways to exercise the control that these central actors are required today. Although this is possible, it is also unlikely.

Incentives seem to be aligned to maintain and protect the central role that the actors analyzed here have achieved. For states, it is easier to control the flow of information when there are central actors with control capacity than when they are absent, there are too many or do not concentrate significant portions of the traffic. The coincidence of those interests with the private interests of some of the most powerful corporations in the world suggest that the current regulatory path is aimed at taking these platforms (and this level of concentration) as a given phenomenon. To the extent that this characteristic of the Internet subsists, and to the extent that diverse actors continue to demand concrete actions from the platforms against threats to democracy perceived as serious, greater transparency regarding the criteria used for moderation seems a reasonable demand. In this context, platforms assume a role that is increasingly similar to that of “public forums”. This conclusion, we argue, is perhaps not only not unavoidable but not desirable.

Mario Bravo 1050, 7º piso (C1175ABT), CABA, Argentina | TE: (54 11) 5199-4500 int 1213 | Fax: (54 11) 4963-1560 | cele@palermo.edu | **www.palermo.edu/cele**